	Informe de Prácticas Profesionales como Opción de Grado	

CREACIÓN DE LA OFICINA DE GOBIERNO DE DATOS Y MEDICIÓN DE LA CALIDAD DE DATOS

PRESENTADO POR:

Jhorman Jair Guetes Miranda

Código:

2016104060

PRESENTADO A:

Juan Carlos Herrera Periñan
Tutor de prácticas

Janier Fabian Moreno Acosta
Jefe inmediato empresa

UNIVERSIDAD DEL MAGDALENA
FACULTAD DE INGENIERÍA
INGENIERÍA DE SISTEMAS

Fecha de entrega: 02/02/2023

Contenido

1. PRESENTACIÓN 3
2. OBJETIVOS Y/O FUNCIONES 4
2.1. OBJETIVO GENERAL: 4
2.2. OBJETIVOS ESPECÍFICOS: 4
3. JUSTIFICACIÓN: 4
4. GENERALIDADES DE LA EMPRESA: 5
5. SITUACIÓN ACTUAL 7
6. BASES TEÓRICAS RELACIONADAS 9
7. DESARROLLO DE ACTIVIDADES: 10
8. CRONOGRAMA: 32
9. CONCLUSIONES Y LÍNEAS FUTURAS 33
10. BIBLIOGRAFÍA 35
ANEXOS 37

1. PRESENTACIÓN

El proyecto por realizar es una propuesta a la empresa Constructora Bolívar S.A. el cual se trata de implementar una oficina de Gobierno de Datos con el fin de cumplir un objetivo estratégico interno de la compañía. El objetivo estratégico mencionado con anterioridad tiene como meta en convertir a la compañía en Data Driven, ¿eso qué significa?, significa que la empresa tomará las decisiones a través de los datos y no por instinto.

Se hizo la implementación de este proyecto dado que puede tener un impacto positivo para la empresa, debido a que la oficina de Gobierno de Datos tiene la misión principal ordenar los datos más importantes de la compañía y medir los mismos, por eso con el apoyo de esta oficina los índices de los datos de la compañía aumentarán dando como resultado datos más confiables a utilizar. Es necesaria estas características que proporciona la oficina, por el hecho de que nos da ventaja frente a la competencia ya que las decisiones tomadas sean más acertadas.

Uno de los motivos principales para la construcción de este proyecto fue el estado de los datos. Esto se detectó debido a una investigación que se realizó, donde los resultados de dicha investigación reflejaron que los datos de la compañía no estaban organizados y además la calidad de estos era muy baja, a raíz de este problema cuando los integrantes de la compañía hacían reportes o dashboard basados en los datos, estos no reflejaban la realidad o plasmaban resultados diferentes para las mismas preguntas de negocio. Debido a esto se llegó a la conclusión de emplear un plan de remediación con el fin de poder sacarle el provecho máximo a los datos y hacer que la compañía pueda tomar las decisiones a través de ellos, al mismo tiempo esto daría como fruto una ventaja en comparación a la competencia.

Para dirigir de manera efectiva y solventar los inconvenientes presentados en los datos era necesario implementar la oficina de Gobierno de Datos, por el hecho de que esta oficina tiene estándares enfocados en dar soluciones a los inconvenientes mencionados anteriormente. En Constructora Bolívar se decidió trabajar bajo el cuerpo de conocimiento llamado DAMA DMBOK, este estándar brinda recomendaciones y menciona cuáles son las buenas prácticas para trabajar con los datos. Dentro de este proyecto se está haciendo la creación de Datawarehouse y varios Data Marts, estos datos se están llevando se están copiando directamente en la nube con herramientas, tales como: SQL Server, Data Factory, Azure Data Lake, Databricks, Azure Synapse y hacer la visualización de los tableros en Power BI.

2. OBJETIVOS Y/O FUNCIONES

2.1. OBJETIVO GENERAL:

Crear la oficina de Gobierno de Datos para generar confianza en los datos y tener mejor toma de decisiones

2.2. OBJETIVOS ESPECÍFICOS:

1. Apoyar en la evaluación de los riesgos y definición de reglas de calidad de datos
2. Documentar todo lo referente a la gestión de problemas concernientes a los datos
3. Participar en los comités de gobierno de datos
4. Hacer el levantamiento de procesos o flujos de los datos, para un buen entendimiento del linaje de los datos.
5. Asegurar y monitorear que las reglas de calidad de datos y procedimientos se cumplan.

3. JUSTIFICACIÓN:

Constructora Bolívar S.A. es una empresa tradicional, la cual busca cambiar su modelo de toma de decisiones. Para lograr este cambio se está apoyando en las tecnologías, dado que a esta situación la empresa se encuentra en una problemática. Esta problemática consiste en que sus datos no están con la calidad suficiente y están de una forma desorganizada. Dicha situación no se soluciona adquiriendo otras herramientas, ya que si no se entiende el problema el mismo aumentaría.

Debido al panorama en el que está envuelta la empresa, el presidente de la compañía junto con sus gerentes tomaron cartas sobre el asunto. La solución pertinente que encontraron es la creación de la oficina de Gobierno de Datos, con el objetivo de organizar, medir y mejorar la calidad de los datos de la compañía. Puesto que con anterioridad se intentó crear un plan de remediación de los datos, sin embargo, el resultado fue distinto al esperado. En otras palabras, este intento de implementación de la gobernanza de los datos y remediación de los de los mismos fracasó ya que no se entendió la situación de la compañía y seguían implantando malas prácticas.

Otro de los puntos por lo que fracasaron tuvo que ver con que no estaban implementando una oficina, sino que intentaron implementar la gobernanza de datos con personas externas de la compañía por un periodo de tiempo bastante corto. Con todas estas enseñanzas adquiridas el presidente de la compañía juntos a sus gerentes decidió llevar a cabo la gobernanza de datos, pero esta vez sí crearían una oficina de Gobierno de Datos con buenas prácticas y además la

compañía ya no trabajaría en silos, de la misma forma las diferentes áreas se comunicarían de la manera apropiada las unas con las otras.

Actualmente, gracias a esa decisión se está llevando a cabo la creación de la oficina Gobierno de Datos, con tal de mejorar la calidad de los datos, pero tomando como referencia al cuerpo de conocimiento el DAMA DMBOK. Esta oficina se creó con el objetivo de no incurrir en el mismo error, el cual era que las áreas trabajan en silos, siendo esta un medio de conexión entre el negocio, los datos y las herramientas que se usan para tener un mejor almacenamiento de los ellos.

4. GENERALIDADES DE LA EMPRESA:

Constructora Bolívar S.A. es la constructora más grande de Colombia, está empresa cuenta con más de 3.000 empleados, por más de 39 años han cumplido los sueños de inversión, desarrollo y vivienda propia de más de 165.000 familias. Esta empresa nació en 1983 bajo el nombre de "Construcciones, Estudios y Proyectos S.A. CEPSA" gracias al respaldo del Grupo Bolívar, con el objetivo y la visión de construir viviendas de calidad para miles de familias colombianas; este objetivo se hizo realidad al consolidarnos como una compañía especializada en el campo inmobiliario. Once años después de la creación de la empresa se le cambió el nombre de la empresa y se le colocó Constructora Bolívar, desde entonces la experiencia y trayectoria se ha especializado en el diseño, la gerencia, construcción y venta de viviendas para todos los estratos sociales.

La compañía tiene cinco valores principales, los cuáles son; Respeto, honestidad, equidad, disciplina y alegría, entusiasmo y buen humor. A pesar de que esta empresa tiene poco tiempo de existencia ha hecho muchas actividades innovadoras y ha asumido retos importantes con el fin de darle a sus clientes unos inmuebles de calidad. En toda esta historia ha obtenido logros importantes los cuales han podido impactar positivamente a miles de personas.

La empresa Constructora Bolívar S.A. tiene sedes en las ciudades más importantes de Colombia y proyectos en toda Colombia, la sede oficial de Constructora Bolívar S.A. se encuentra en la ciudad de Bogotá y la dirección de dicha sede es **Calle 134 #72 - 31**, sin tener en cuenta que su segunda sede más importante es en la ciudad de Cali**.**

Dentro de la compañía trabajan muchas personas, algunas de ellas de manera directa y otras de forma indirectas, por tal razón es necesario entidad tenga una buena organización. Debido a esto, el organigrama de la organización es bastante extenso, por esto mismo cuenta con un presidente, dicho presidente tiene a cargo trece gerencias, donde se toman las decisiones estratégicas y más importante de la empresa. El número de empleados que existen en la organización de superior a los tres mil. Actualmente una de esa gerencia es la Transformación (tecnológica) donde hay cinco direcciones y en una de esas direcciones es la de Big Data &

Analytics, dentro de Big Data & Analytics se encuentran áreas tales como: Ingeniería de Datos, Analítica Avanzada, Business Intelligence y Gobierno de Datos. Cada una de estas áreas son las encargadas manipular los datos y que empresa puede lograr los objetivos internos para así de este manera convertir a la compañía en Data Driven, cada área ayudando desde su rol, responsabilidades y habilidades.

En esta compañía hay una cultura muy marcada, esto se debe al código de ética que se maneja, también por la visión y la misión de esta. Esta compañía tiene como **misión** generar valor económico, social y ambiental en forma sostenible para beneficio de la comunidad y de nuestros accionistas, colaboradores y clientes. Para lograrlo, contamos con un equipo humano que se caracteriza por sus destrezas en: Conocimiento de Nuestro Negocio y de Nuestros Clientes Manejo del Riesgo Relaciones Emocionales Investigación, Innovación y Tecnología Sinergia Empresarial.

Esta misión está muy alineada a la **visión** la cuál es ser una compañía con presencia Nacional orientada al cliente, que ofrece productos de calidad y que protegen el medio ambiente; participando activamente en el mercado de vivienda y siendo líder en vivienda de interés social, dando un importante apoyo a nuestros compradores para generar comunidades sostenibles. Teniendo además presencia importante en el mercado de desarrollos turísticos, segunda vivienda, de comercio y oficinas.

Constructora Bolívar pertenece a uno de los grupos económicos más sólidos del país de Colombia, La Familia Bolívar, de la que hacen parte el Banco Davivienda, Seguros Bolívar, Seguros Libertador, para mencionar sólo tres de las más importantes. Este grupo intenta alienar a todas las compañías que tiene con un código de ética, el cual es un factor determinante en la cultura de las empresas del Grupo Bolívar y es altamente valorada por nuestros funcionarios, clientes, accionistas y público en general. Es considerada como una parte fundamental de nuestra gestión, dado que hace parte de la conciencia individual y colectiva. Lo anterior se hace evidente en la preferencia del público por nuestras marcas, productos y servicios, y en la reputación e imagen construidas por años sobre dicho fundamento, lo que representa un activo estratégico de las compañías y, además, constituye un respaldo esencial en momentos de eventual dificultad. Este código de ética hace parte integral del Sistema de Buen Gobierno Corporativo que han adoptado las empresas del Grupo Bolívar, el cual pretende, entre otros aspectos, que la confianza y valor percibidos por funcionarios, clientes, accionistas y demás grupos de interés aumenten y seamos reconocidos por la mayor percepción de transparencia en nuestra gestión. En el código se enuncian algunas pautas básicas de conducta que deben ser cumplidas por los miembros de las Juntas Directivas, presidentes, directivos, funcionarios e Intermediarios. Pero es importante resaltar que se hace necesario siempre acudir al buen criterio, la responsabilidad y la prudencia a la hora de valorar una situación de posible conflicto ético. Por último, los invito a que nos familiaricemos también con nuestros principios y valores, las leyes y regulaciones aplicables,

las políticas internas, y actuemos siempre en consecuencia, incluyendo la solicitud de consejo a las instancias pertinentes si hay duda sobre el curso apropiado de acción, o recurriendo a los jefes inmediatos ante una situación de conflicto de interés entre lo personal y lo empresarial. Estas fueron las palabras dichas por el sr. Miguel Cortés Kotal es el presidente de todo el grupo Bolívar.

5. SITUACIÓN ACTUAL

Durante el proceso de prácticas profesionales la empresa estaba desarrollando un modelo de negocio diferente al que quiere desarrollar, por tal razón se vió en la necesidad de implementar una oficina de Gobierno de Datos, con el objetivo de poner en marcha un plan de negocio que le ayude a cumplir con unos objetivos estratégicos. Esto se puede reducir a que la empresa tenía muchos inconvenientes a la hora de confiar en sus datos porque los mismos no reflejaban la realidad del negocio. Muchas veces para comunicarse con sus clientes, la compañía se veía obligada comprar los datos de contractibilidad de ellos. Eso tiene como desventaja que tienen la necesidad de pagar a un tercero por los datos de sus propios clientes y que la operación se vuelve menos eficiente para los asesores de la compañía o el call center encargado de comunicarse con los clientes. Esto trae como consecuencias el gasto de forma ineficiente los recursos de la compañía tales como: Recurso económico y el recurso del tiempo de operación de los asesores, eso se traduce a pérdidas en horas nominal y afecta a el presupuesto de la empresa.

Por todo esto lo que estaba pasando en la compañía era inevitable llegar a conclusiones acertadas, tales como: Si los datos de los clientes que son el pilar de cualquier compañía lo tenemos mal y no son confiable, pues, los datos de toda la compañía en general están mal. A estas conclusiones llegaron también cuando intentaban hacer un reporte o informe importante, se daban cuenta que los datos de un reporte no eran igual a los de otros a pesar de que intentaban solucionar o responder las mismas preguntas.

En la compañía también existe una situación la cual puede ser uno de los causantes de los problemas mencionados anteriormente. Actualmente en la organización se trabaja con un ERP, este genera muchas bases de datos, evidentemente es un problema porque la información se duplica o en su defecto vienen muchos datos nulos, ya que el sistema no está tan optimizado como debería. En este ERP se puede notar que los datos de buena calidad son muy escasos. Esto ya se ha ido tratando tiempos atrás de varias maneras, una de las soluciones fue crear un nuevo sistema de información. El nuevo sistema es un CRM, ya que en el actual ERP intentaba hacer las funciones de CRM también, debido a que los datos estaban creciendo cada día más y las consultas para buscar un cliente o simplemente las finanzas de la empresa eran inmanejables decidieron crear el nuevo sistema. Esta solución trajo consigo muchas puntos positivos y puntos negativos, dado que no se tuvo en cuenta una arquitectura en

específico, y en este momento la sincronización entre ambos sistemas es muy difícil. A esto hay que sumarle el hecho de que ninguno de los sistemas que tiene la compañía cuenta con un diccionario de datos donde estén mapeados los atributos más importantes o que más se usen, para encontrar algún dato en específico es necesario sumergirse en la base datos o preguntarle a una persona que cuente con esa información y experiencia en la compañía.

A estas situaciones comentadas y con la visión que tiene la empresa de guiarse de los datos, pues, hace que busque soluciones a estas situaciones, una de las primeras soluciones a las que llegaron fue la creación de la oficina con el carácter y el conocimiento para gobernar los datos. A pesar de que esta oficina no tiene mucho tiempo en la compañía ya se ha ganado la confianza y esperanza de esta debido a el trabajo y el entendimiento que ha tenido del negocio. Otras de las soluciones que están implementando es contratar a terceros que ayuden a mejorar la situación actual de los datos de la compañía, la empresa consultora Accenture. Esta empresa consultora está proporcionando personal de arquitectura, ingeniería de datos y especialistas en gobierno de datos para que la dirección de Big Data & Analytics, con el fin de esta área esté alineada y haga uso de buenas prácticas tales como mejorar la arquitectura de datos.

Una de las tareas que tiene la dirección con la consultora es usar las herramientas que se necesiten, por tal razón se está haciendo uso de herramientas de la nube con el fin de crear un Data Lake House, donde se está llevando a cabo por datamarts. En el día de hoy se está haciendo el mapping de los campos usados en las ventas y extrayendo los datos de las fuentes On-Premises y subiendo en la nube para tener unos datos bien ordenados y listos para el uso.

Gobierno de Datos está ayudando en esta situación de la siguiente manera, es la encargada de supervisar que los datos que se están mapeando sean lo correctos, también está involucrada en el levantamiento de información ya que es una oficina con mucha experiencia en el levantamiento de información y eso es necesario para tener resultados rápidos en el levantamiento, otras de las acciones que hace Gobierno de Datos es ayudar en el mapping de los atributos y levantar la mano si ve que la entidad relación del datamart no refleja la realidad del negocio, ya que, si no refleja la realidad del negocio este no le podrá sacar el provecho suficiente.

Una de las visiones futuras que tiene Gobierno de Datos es que cuando se terminen implementar los datamart, la oficina de Gobierno de Datos tendrá el poder de decidir quién será la personas o el área podrá usar la información. Esto se da porque la oficina de Gobierno de Datos es la encargada de analizar las peticiones de los clientes internos sobre los datos que desean visualizar. Por esta razón analiza si la persona que hace petición tiene motivos fehacientes para recibir la información solicitada, si dicha persona o área tiene la necesidad de ver la información la oficina de Gobierno de Datos le dejará acceder la información que solicita.

Todo esto es lo que se tiene visualizado, pero, realmente en este momento la situación no es la más fácil de resolver, debido a muchos factores, hay que hacer tablas de homologación, es necesario hacer la ingesta de la información a la nube y como es mucha información y muchas bases de datos esto puede demorar. Otras de las dificultades más grandes es que como el ERP genera muchas bases de datos y no tienen un estándar entre sí, esto hace que las homologaciones no sean lo más sencillo posible, a esta situación hay que sumarle que el CRM que tiene su propia estructura y naturaleza.

6. BASES TEÓRICAS RELACIONADAS

Las bases teóricas adquiridas en la etapa universitaria fueron fundamentales para desempeño de las labores y responsabilidades otorgadas en las prácticas profesionales. Empezando con materias que impartidas en los primeros semestres tales como; **Pensamiento de sistemas** y **Algoritmo y Programación**. Estás materias ayudaron en el trabajo del practicante porque el rol que desempeña. En las actividades diarias del practicante debía buscar soluciones a problemas y en la materia de Pensamiento de Sistemas impartían ese conocimiento y métodos de como buscar soluciones de manera específicas o general aplicando el pensamiento sistémico. Otro de los pilares aprendido en el proceso universitario fue la asignatura Algoritmo y Programación, esta fue la bases para entender la lógica de programación, debido a que una de las actividades realizadas por el practicante era programar en Databricks con el lenguaje de programación Python, SQL y la librería de Python llamada PySpark, a pesar de que todo el conocimiento de programación adquirido en la universidad, la materia Algoritmo y Programación fue el cimiento para el desarrollo, entendimiento y conocimientos adquiridos después de la etapa universitaria.

Otras materias importantes cursadas en la carrera e importantes para el desarrollo fueron **Estructura de datos I**, **Estructura de datos II** y **Compiladores**, como en la materia de Algoritmo y Programación el conocimiento fue básico, solo con ese conocimiento no habría podido llevar a cabo las actividades profesionales, gracias a las materias de Estructura de Datos quedó claro qué manera se optimiza un código y cuál es la forma como se debe emplear las estructuras tales como: Pilas, cola, etc., esto sin contar que en la Compiladores se impartió las bases del lenguaje de programación Python. En esta materia no se hizo programas con la librería PySpark, pero, el aprendizaje obtenido en esa materia ayudó para que la curva de aprendizaje fuera mucho más rápida.

Está claro que las materias mencionadas anteriormente ayudaron a tener un buen desempeño y a cumplir con las tareas otorgadas. Sin embargo, como las practicas fueron desarrolladas en el área de Big Data & Analytics, lo que significa que era necesario tener contacto todos los días con bases de datos y querys de SQL. Por tal razón, fue necesario investigar de estos

temas mientras se cursaba las prácticas profesionales, sin embargo, eran dudas muy puntuales debido a que en la formación universitaria hubo contactos con consultas de bases datos de datos. En materias como **Bases de Datos**, **Seminario I, Seminario II** y **Seminario III** impartió conocimientos de bases de datos, desde lo dimensional hasta las consultas realizadas para mostrar la información que se necesitaba. A todo esto, hay que añadirle los plazos y las responsabilidades que se debía tener para llevar a cabo las tareas asignadas.

Todas las temáticas y materias mencionadas fueron las bases técnicas, sin embargo, eso no quiere decir que solo fueron las materias utilizadas, el rol que tenía el practicante en la compañía era diversas exposiciones y también entendimiento del negocio, o sea, la forma como la empresa tiene constituida su organigrama y de qué manera funciona. En esta fase más teórica fueron necesarias materias teóricas de la carrera. En **Expresión Oral**, se adquirió tácticas de cómo expresarse las ideas de la forma correcta. También, el conocimiento impartido enseñó de cómo expresarse para que la audiencia esté pendiente de los temas que se tratan. En esta materia también se enseña cómo no tener un tono de voz monótono, sino que es necesario a la hora de exponer variar el tono de voz, hacer pausas para que las personas no sientan que es aburrido lo que está expresando.

A pesar de esto, como los temas tratados eran como funcionaban las diferentes área, direcciones y gerencia de la compañía era necesario tener un conocimiento previo de cómo funcionan los negocios y cuáles son los diferentes métodos. Estos conceptos y términos fueron compartidos en materias tales como: **DOTI** y **Ética Profesional**. En la materia como DOTI específicamente fue donde se trató de qué manera funcionan los negocios o cuál es la naturaleza de los mismo. Gracias a esto se realizó aportes en la compañía que resaltaron todo este conocimiento y entendimiento de los temas tratados.

En la materia Ética Profesional fue donde se compartió el conocimiento de cómo se comporta un profesional y de qué manera debe actuar para el beneficio de la empresa. Gracias a este tipo de conocimiento se logró identificar si los comportamientos que hacía el practicante eran políticas correctas y cómo debía de actuar por el bienestar de la profesión, por el bienestar de mis compañeros y por el bienestar de la compañía.

7. DESARROLLO DE ACTIVIDADES:

Los primeros días de prácticas profesionales los temas tratados eran muy confusos, no se entendía mucho cuál era el objetivo del área o cuál era el aporte que esta debía hacer con respecto a la compañía. En estos días el practicante tuvo participaciones menores, ya que, en el segundo día todo el equipo de Gobierno de Datos destinó ese día a trabajar en las reglas de calidad e identificar los atributos más importantes del dominio Registro de Clientes. Este fue el primer dominio que se iba a activar en el área. En ese día estaba todo el equipo en una

oficina haciendo lluvia de ideas de los atributos que pertenecían a dicho dominio de datos y a pesar de que el practicante no tenía conocimiento en qué eran las dimensiones de calidad o cuáles eran, la lógica que tenía desarrollada gracias en la programación le ayudaron a no sentirme perdido y poder hacer pequeños aportes a la actividad que se estaba desarrollando. En esta actividad se estaba adelantando las reglas de calidad para posteriormente socializarlas con las personas funcionales y así llevar un trabajo adelantado con respecto a la medición de los índices de calidad de datos.

Aunque, la socialización con el equipo funcional no era de forma inmediata, una de las buenas prácticas de Gobierno de Datos es tener adelantado el trabajo con los temas más importantes definidos antes de hacer las reuniones de socialización. En las primeras semanas el practicante no era capaz de dimensionar el alcance como tal del área. Por tal razón, se vio en la necesidad comenzar a investigar sobre los temas y los términos que no entendía, estas fueron las acciones que realizó para comprender cuáles eran las dimensiones de calidad y cuáles eran los objetivos de cada una, eso sin mencionar que siempre contó con los compañeros de trabajos, los cuales estaban en disposición de ayudar en cualquiera de las dudas que se le presentan al practicante.

Las primeras semanas el practicante siempre obtiene inducciones varios tipos. Una de las inducciones es con el compañero de equipo, en esta inducción se le hace la explicación de cómo trabaja el equipo y el rol que debe tomar el practicante dentro de la oficina. En esta inducción también se menciona qué se espera del practicante y cuáles son las actividades que realizará. Otra de las inducciones es para conocer la cultura de la empresa y la manera como trabaja, explicándole todas las gerencias y la visión y misión de cada una de ella.

Hay una inducción que le dicta el jefe inmediato, esta tiene como objetivo que el practicante sepa el organigrama de la dirección, también se menciona como es que cada área de la dirección debe complementarse o cuál es la responsabilidad de cada una. También se le muestra cuántas áreas está compuesta la dirección, a cuál se hace parte y cuál es el aporte de Gobierno de Datos dentro de la dirección y en la compañía.

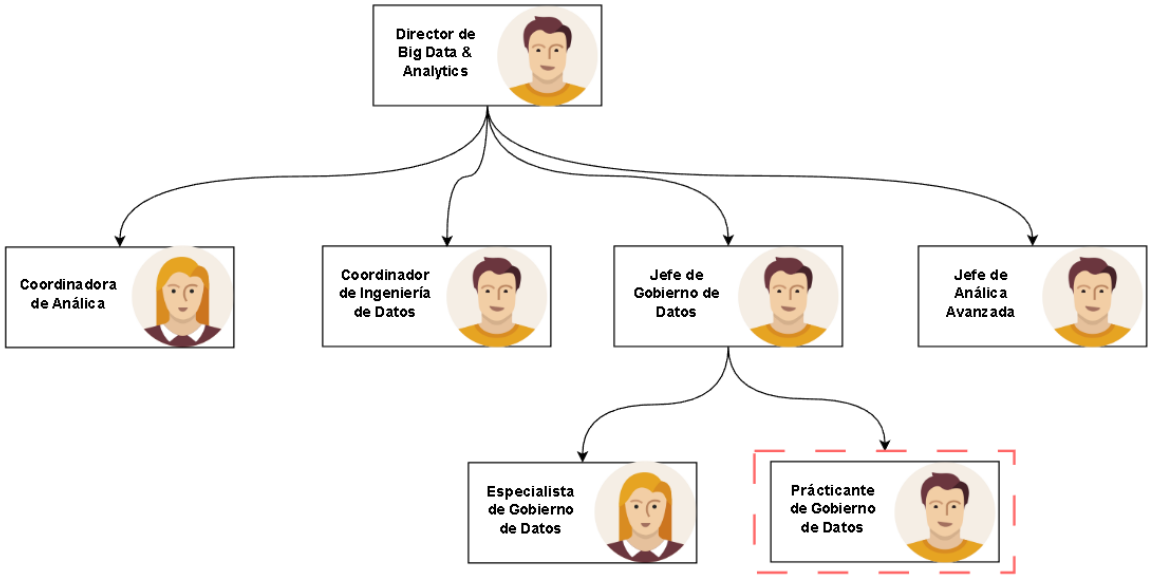

Ilustración 1: Organigrama de Big Data & Analytics

Todas estas explicaciones son necesaria, ya que, es importante conocer cuál es la dirección donde se está y el practicante debe tener claro. Esto es importante porque la oficina interactúa con personas de diferentes gerencias y saber en qué rama se está ubicado en el organigrama es más fácil presentarse o explicarle a las personas que se quieren entrevistar.

El conocimiento que se imparte en la inducción es la explicación de la oficina de Gobierno de Datos, este se compone por un roadmap de trabajo donde se describen los pasos a seguir de la oficina y de qué manera se va a implementar el modelo operativo. Acá es donde se le explica al practicante las tres maneras globales de implementar la oficina de Gobierno de Datos las cuales son: Centralizado, federado y descentralizado.

En la compañía en conjunto de un consultor llamado Accenture se hizo una evaluación donde se llegó a la conclusión que para la compañía la opción que más le sirve y se adapta más a ella es la federada por sus diferentes características, pero, ¿qué significa operar de forma federada?, significa que toda la compañía debe tener sentido de pertenencia con los datos y que deben saber la importancia de tener los datos correctos, deben tener en la mente que todos somos los encargados de velar por los datos, que estén actualizado y lo más importante que sean confiable para poder usarlos a nuestro favor. Sin embargo, para generar este tipo de pensamientos en las diferentes personas y áreas de la compañía es necesario hacer un cambio de cultura, por tal razón, la oficina de Gobierno de Datos se debía hacer reuniones de manera periódica junto con el grupo de Gestión del Cambio, se hacía con el objetivo de que guiara a

la oficina en este proceso, así de esta manera construir una cultura consciente de los datos y la importancia de estos.
En esta inducción se le dice al practicante las conclusiones de la investigación que se realizó, se pudo observar que la compañía no tenía las características para que se implementara las estructuras centralizada ni las descentralizada, ya que, para implementar la primera estructura mencionada era necesario tener un equipo muy grande en la oficina de Gobierno de Datos orquestando todo lo que pase en la compañía con respecto a los datos, eso traería puntos negativos tales como que la cultura de la compañía y la visión que tiene sobre los datos no cambiaría y uno de los propósitos difundir esta importancia. Por otra parte, la forma descentralizada había necesidad de tener una oficina de Gobierno de Datos en cada área de la compañía, esto no era viable por dos razones, la primera razón era el capital que debía tener la compañía para ejecutar dicha manera de trabajo, y la otra es que las áreas trabajan de manera separada, cosa que se quiere erradicar para tener una compañía con comunicación entre las áreas. También no era viable porque la madurez o la conciencia de los datos en la compañía ya debía de estar elevada, en este caso no aplicaba porque muchas personas no tenían idea de qué manera la compañía iba a ser beneficiada por tener unos datos correctos y ordenados.

En la siguiente imagen hay una pequeña explicación de cuáles son los pros y los contras de cada uno de este tipo de estructura. En la imagen a continuación se puede ver de una manera más gráfica estas estructuras con sus respectivas ventajas y desventajas.

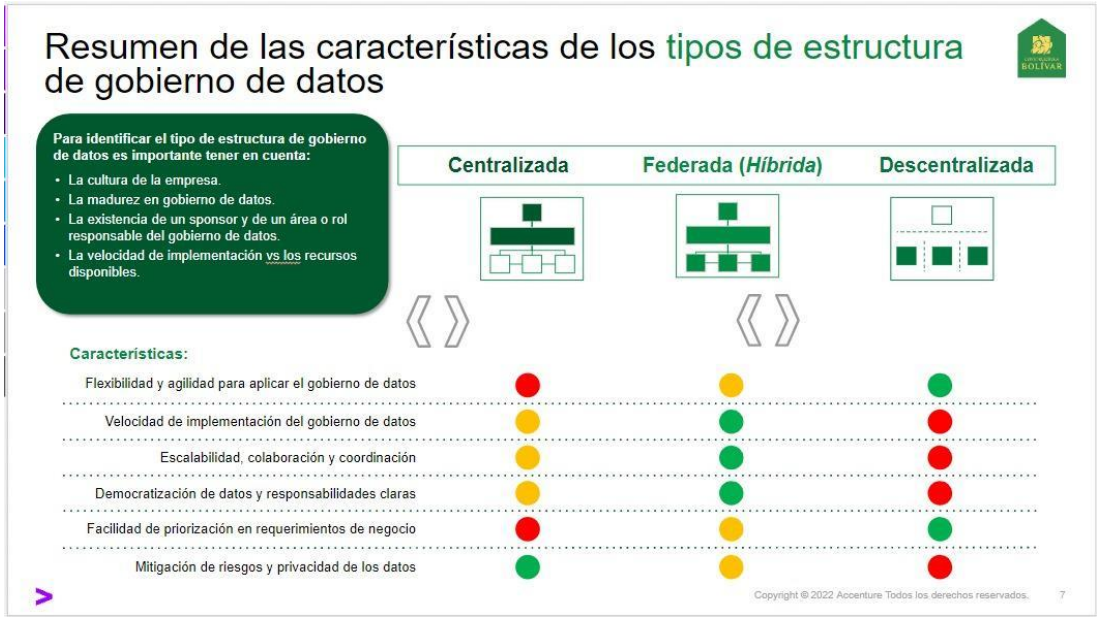

Ilustración 2: Tipos de estructuras de Gobierno de Datos

En las inducciones se hacen explicaciones generales donde le explican al practicante cuál es la visión de la visión de la oficina. Teniendo el conocimiento de la inducción el practicante debe tener claro el papel que debe ejercer y por qué es necesario un practicante con conocimiento en bases de datos y negocios. Una de las acciones que realizó el practicante fue leer el libro que dictaba el estándar de la oficina, el estándar es DMBOK DAMA. Aunque, este libro no es necesariamente corto, sin embargo, para serle útil a los compañeros de trabajos y a la oficina. Uno de los impedimentos con los que se topó a la hora de leerlo es la manera como estaba redactado el libro, ya que, este es bastante técnico y medio complicado de entender. Debido a esto, intentó encontrar una comunidad en internet que le ayudara con el entendimiento de la información que tiene el libro. Por suerte, encontró varias comunidades en YouTube, pero, la que principalmente llamó su atención fue la de **Data Engineering Latam**, en este canal en específico ofrece webinar de resumen de los diferentes capítulos del libro. En estos webinar una persona lideraba la sesión y de qué manera lo interpretó, sin embargo, había más personas conectaban que decían si estaban de acuerdo y cuáles fueron sus experiencias con respecto a ese capítulo en específico. Este tipo de información le ayudó a entender mucho más los temas que estaba conviviendo en las prácticas.

Ya entendiendo la necesidad de la oficina el practicante tuvo que colocar en práctica el conocimiento obtenido en la universidad, fue aquí donde comenzó a interesarse por la parte de las bases de datos, ya que, había visualizado de qué manera estás están ligadas al Gobierno de Datos, en estos días el practicante estuvo en unas sesiones de entendimiento de cómo estaban las bases de datos y los datos de la compañía, luego por su iniciativa y análisis pudo observar que habían cierta anomalías dentro de las bases de datos. Cada una de estas anomalías fueron expuestas al equipo de Gobierno de Datos. Esta actitud fue la que dio un lugar importante en el equipo, debido a que el equipo necesitaba una persona con conocimiento en bases de datos, estas actividades debían ser realizadas por el rol de profesional o analista de calidad de datos, sin embargo, como la oficina se estaba iniciando no existía este rol en específico.

Es rol de analista de calidad de datos fue tomada por el practicante, fui a ahí donde comenzó a buscar la información y los atributos de las diferentes bases de datos. También se hacía con el de observar que los datos cuentan con la calidad necesaria. Debido a esas investigaciones realizadas por el practicante encontró muchas anomalías en los datos, anomalías tales como: Una tabla tiene dos campos que querían guardar los mismos datos, o sea, campos repetidos, cosas como que muchas columnas donde se supone que debería haber datos estaban totalmente vacíos, o tablas que literalmente estaban vacías. Todas estas situaciones fueron presentadas al grupo de trabajo y discutido con ellos, en estas reuniones se plantearon las diferentes formas de mejorar los datos o simplemente darnos una idea de cómo iban a salir los resultados de los índices de calidad que íbamos a implementar en un futuro. Todas esas

inconsistencias fueron socializadas al jefe directo del practicante, donde este habló con el director del área y así escaló el problema encontrado.

En la siguiente imagen se mostrará una tabla llamada TercerosCiudad la cuál guarda datos vitales para el negocio están totalmente vacíos en esa base datos, esto también pasó por la forma como están organizada la empresa en las bases de datos.

```
SELECT [TrCTercero]
      ,[TrCIdCiudad]
      ,[TrCContacto]
      ,[TrCDireccion]
      ,[TrCCiudad]
      ,[TrCTelefono]
  FROM [SINCOSQL].[SincoFideicomisoFontibonHB].[dbo].[TercerosCiudad]
GO
```

100 %

Results | Messages

TrCTercero	TrCIdCiudad	TrCContacto	TrCDireccion	TrCCiudad	TrCTelefono

Ilustración 3: Evidencia de las inconsistencias en las bases de datos

Fue para este mismo tiempo donde comenzó a entender la importancia de los artefactos de Gobierno de Datos y cómo mapear cada uno de los procesos. En esta oficina se usan aproximadamente siete artefactos, cada uno con un propósito en específico. En este tiempo el practicante tuvo la experiencia y entendimiento suficiente de la oficina para poder mejorar los artefactos y también crear unos artefactos nuevos. Los artefactos más importantes eran los siguientes: Glosario de negocio, diccionario de datos, diccionario de atributos, mapeo de ciudadanos e inventario de reglas de calidad.

Como cada uno de los artefactos tiene un objetivo diferente, ya que, esto se hacía con el objetivo de tener una constancia sólida del levantamiento de información que se hacía en la oficina. ¿Cuál era el objetivo de cada artefacto?, el que más se usaba era el glosario de negocio, este artefacto tenía como finalidad tener todos los conceptos de la compañía en un solo lugar, esto se hace porque muchas veces los conceptos de la compañía se tergiversan o

algunas personas tienen unas definiciones que no corresponden a lo que el negocio se refiere. Por esta razón debe haber una oficina como la es Gobierno de Datos que busque la unificación de conceptos, una oficina que orqueste este tipo de tareas. El proceso para llevar a cabo esta labor era el siguiente, como en todos los demás artefactos, primero la oficina de Gobierno de Datos en busca de la activación de un dominio o en busca de organizar un conjunto de datos en específico debía ayudar hacer entrevistas a las personas funcionales, personas que trataban los datos y los conceptos día a día. Después de una levantamiento y entendimiento del negocio la labor de Gobierno de Datos era darle un concepto claro a las palabras que las personas entrevistadas decían, esta era la tarea inicial, ya teniendo esas propuestas, se hacían diferentes reuniones con las personas con más conocimiento del área y con más influencia con el objetivo de hacer un concepto estándar y que esté alineado a cómo se maneja en la compañía.

Glosario de Negocio

IdDominio	Dominio de datos	Nombre	Descripción	Dominio responsable del concepto compartido	Fuente del concepto	Certificado por Negocio SI / NO
MD001	Registro del cliente	Cliente	Persona natural o jurídica que confirma mediante la consignación el valor de la separación de un bien inmueble que planifica o esta construyendo constructora Bolívar, que genera un contrato y unas obligaciones contractuales.			SI

Ilustración 4: Artefacto Glosario de Negocio

En la imagen podemos observar un ejemplo, el concepto es cliente, la definición ya fue avalada por los roles del ese dominio que en este caso es Registro de Cliente, un concepto que se tenía dudas y que personas dependiendo de a quién se le preguntara definían un tanto diferente, pero, con el levantamiento y las reuniones con personas importantes de diferentes gerencias se pudo llegar a esa definición, por esa razón es la definición estándar por el negocio.

Otro artefacto importante y en el que el practicante tenía la responsabilidad total de él es diccionario de datos, este tiene como foco principal tener mapeados los atributos más importantes de la compañía, con el fin de saber dónde se encuentran dichos atributos y dónde podemos encontrarlo para hacer informes, tableros de BI, etc. Este es el primer paso para hacer una democratización de los datos y tener una buena disponibilización de los mismos. Como el practicante era el encargado de este artefacto se dio cuenta de ciertas limitaciones, ya que era un poco confuso trabajar con él y saber cuáles eran los atributos que faltaban o cuáles ya estaban mapeados, esto era así porque muchos atributos estaban en diferentes fuentes de información. Este artefacto se comunicaba con dos más, el cuál es diccionario de atributos e inventario de atributos, este último fue creado por el practicante, lo creó para llevar a cabo un mejor manejo a la hora de mapear los atributos.

Diccionario de Datos

Código del atributo	Dominio de datos	Nombre del atributo	Fuente / Aplicación / Sistema	Descripción de la Fuente / Aplicación / Sistema	On Premise / Cloud	Fuente autorizada	Servidor
MD001RCL001	Registro del clientes	Tipo de documento de identificación	SerK		Cloud	Si	CRM - AZURE
MD001RCL001	Registro del clientes	Tipo de documento de identificación	Sinco		On Premise	Si	SINCOSQL

Ilustración 5: Artefacto Diccionario de Datos

Acá podemos ver que el diccionario de datos dice la metadata de cada atributos, en qué base datos está, cuál es el servidor, tamaño, etc. esto se complementa con el diccionario de atributos porque este artefacto era el encargado de reflejar la definición de cada atributo. Este último artefacto se hacía con anterioridad, con el fin de tener claro a qué hacía referencia cada atributo y cuando se fuese a mapear se supiera que ese atributo si estuviesen acorde con las definiciones marcadas anteriormente. Para que se entienda, un el diccionario de datos eran donde se mapeaban y se llevaba la información de la metadata de cada atributo, pero con el diccionario de atributo lo que se llevaba era la definición de cada atributo para saber cien por ciento a qué hacía referencia. Estos dos artefactos trabajaban bien en conjunto, sin embargo, a veces era tedioso saber cuál atributo ya estaba, a cuál había que socializar con los DBA, cuales no fueron encontrados. Por esa razón el practicante creó un nuevo artefacto, uno donde se puede ver la información de una manera más visual, lo cual es lo que se necesitaba. Fue de esta manera creó el artefacto llamado inventario de atributos, este tiene el objetivo de saber cuáles son los sistemas de información que ya se mapearon, cuáles son los atributos que no existen en alguna de ella, etc.

N°	ID Atributo	Atributos	Campos identificados	CDE's	Diccionario de datos			Diccionario de atributos	Reglas de calidad	Lista de referencia
					SerK	Sinco	CRMWeb			
1	MD001	Tipo de documento	Es atributo	es CDE	Sí existe	Sí existe	Sin gestionar	Está en el diccionario de atributo	Tiene regla	Requiere pero no tiene lista de referencia definida
2	MD002	*Núemro de documento*	Es atributo	es CDE	Sí existe	Sí existe	Sin gestionar	Está en el diccionario de atributo	Tiene regla	No necesita lista de referencia
3	MD003	*Nombre completo*	Es atributo	es CDE	Validar con el BDA	Sí existe	Sin gestionar	Está en el diccionario de atributo	Tiene regla	No necesita lista de referencia
4	MD004	*Profesión*	Es atributo	es CDE	No existe	Sí existe	Sin gestionar	Está en el diccionario de atributo	No tiene regla	Requiere pero no tiene lista de referencia definida
		Totales	0	100%	0	0	4	0	1	0
			4		2	4	0	4	3	2
			0		1	0	0	0	0	2
					1	0	0			

Ilustración 6: Artefacto Inventario de Atributos

Ya con este artefacto se tiene claro en la situación que se encuentra cada atributo, cuál es su estado en las diferentes fuentes y sus datos más importantes, como si es CDE o si tiene una lista de referencia. Gracias a esto se lleva un control y se sabe cuál es el porcentaje de avance, por fuente, por atributo, etc.

Ya teniendo mapeados los atributos es necesario asignarle las dimensiones de calidad a cada atributo, en esta oficina estábamos trabajando con cuatro dimensiones, pero, no eran todas las que se querían manejar. Las dimensiones activas o que se estaban trabajando para medir la calidad de datos de los atributos identificados en los dominios son los siguientes:

Completitud, unicidad, validez y razonabilidad. Todas estas eran las dimensiones que se estaban evaluando, sin embargo, faltaban varias dimensiones por evaluar, pero el objetivo con esas dimensiones faltantes era evaluarlas a futuro. Las dimensiones que no se están evaluando actualmente y que se tiene como un objetivo a futuro son: Consistencia y exactitud. Para desarrollar este artefacto nos reunimos ya sea de manera presencial y virtual y discutimos a qué se acoplada más cada atributo y de qué manera asignarles las reglas de calidad para que no diéramos pie a que se siguieran guardando los datos de manera errónea, en este artefacto también llevamos un inventario de reglas de calidad, con el fin de reutilizar el trabajo. Las reglas se veían de esta manera.

Reglas de Calidad de Datos por CDE

Código del atributo crítico	Dominio de datos	Dimensión de calidad de datos	Definición regla estándar de calidad de datos	Descripción de la regla para el atributo crítico (especificar valores o lógica, si aplica)
2	Registro del Cliente	Completitud	El campo no puede ser nulo	
	Registro del Cliente	Unicidad	Combinado con el Tipo de identificación debe ser único	
	Registro del Cliente	Validez	**R1**: Las reglas lógicas para los formatos de acuerdo al tipo de identificación son: Cédula de Extranjería (CE): - Rango de 1.000 hasta 999.999. Cédula de Ciudadanía (CC): - Rango de 100 a 99.999.999 (antiguas) y de 1.000.000.000 hasta 1.999.999.999 (nuevas) - Para las cédulas que tienen 8 dígitos o menos, se debe validar contra los lineamientos de la Registraduría Nacional Tarjeta de Identidad (TI): - Rango de 1.000.000.000 hasta 1.999.999.999 (Nuevas) Registro Civil (RC): - Rango de 1.000.000.000 hasta 1.999.999.999 (nuevos) Número de Identificación Tributaria (NIT): - Rango de 800.000.000 hasta 999.999.999 NIT de Extranjería (NTE): - Rango de 600.000.000 hasta 799.999.999 NUIP: - Tiene 10 dígitos y empieza por el rango de mil millones 1.000.000.000 **R2**: Para persona jurídica, el campo debe contener al menos 9 dígitos para que se considere completo	

Ilustración 7: Artefacto Reglas de Calidad de Datos

Acá podemos observar de qué manera trabajamos las reglas de calidad, en este ejemplo en específico podemos ver que este atributo hay tres dimensiones de calidad para evaluar, todo esto después de esto se programaba en Databricks, para materializar las reglas ya planteadas. También podemos ver que cada regla tiene un objetivo en específico.

Repositorio Reglas Estándar de Calidad de Datos

RC-0001	Completitud	Mandatorio	El dato no debe ser nulo o vacío
RC-0002	Unicidad	Duplicado	El dato debe ser único
RC-0003	Unicidad	Máximo repeticiones	El dato solo puede estar repetido una cantidad máxima de veces
RC-0004	Validez	Fecha	El dato debe estar en un formato de fecha válido
RC-0005	Validez	Numérico	El dato solo puede tener números
RC-0006	Validez	Alfabético	El dato solo puede tener letras
RC-0007	Validez	Caracteres excluidos	El dato no debe tener los caracteres o palabras que se hayan definido como excluidos
RC-0008	Validez	Estándar de industria	El dato debe corresponder con el formato definido de acuerdo con estándar de industria
RC-0009	Validez	Binario	El dato solo puede uno de dos valores posibles
RC-0010	Validez	Lista definida	El dato debe estar dentro de los valores de la lista de referencia que le corresponda

Ilustración 8: Repositorio de Reglas de Calidad de Datos

De esta manera llevábamos una lista de reglas estándares, para poder reutilizar más adelantes porque hay atributos que se repiten en diferentes dominios y atributos, debido a que hay atributos que se comportan de una manera simular, de esta manera optimizamos el trabajo ya realizado con anterioridad.

Para llevar la información de una forma bien estructurada y correcta del levantamiento de información era necesario saber cuáles eran las personas que se entrevistaron. De estas personas debíamos saber la información general de ellos, cosas tales como: Dominio al que pertenece, nombre de la persona, cargo, gerencia a la que pertenece, tipo de ciudadano, nivel de influencia, nivel de interés, correo electrónico, etc. Toda esta información se recopila para saber si la persona entrevistada cuenta con una serie de características, las cuales son importante para asignarle algún rol dentro de la oficina de Gobierno de Datos, en esta oficina se trabajan con varios roles. Dentro la oficina hay varios roles y fuera de la oficina también, como la oficina se ciñe de una estructura federada. Fuera de la oficina se encuentran los siguientes roles: dueño del dominio, administrador y custodios. Para eso es importante tener claros las características de cada persona, para saber cuál de estos puede cumplir de manera correcta estos roles, generalmente por cada dominio debe tener un solo dueño del dominio, de los otros dos roles puede haber varios, ya sean de diferentes gerencias o de la misma, depende del tipo de dominio.

Mapa de Ciudadanos de Datos

No.	Dominio de datos	Nombres y apellidos	Gerencia	Área	Cargo	Tipo de ciudadano de datos	Nivel de influencia	Nivel de interés	Valor influencia	Valor interes	Correo electrónico
1	Registro de clientes	Jose Daniel Ramirez López	Comercial	Mercadeo	Coordinador	Productor de datos	Alto	Muy alto	4	5	juan.dani.ramirez@constructorabolivar.com
2	Registro de clientes	Maria Esmeralda Gonzalez Pérez	Tramites	Recaudos	COORDINADOR I CALIDAD DEL SERVICIO • CAPITAL SOCIAL - CENTRO DE EXPERIENCIA	Consumidor de datos	Neutro	Bajo	3	2	maria.esme.gonza@constructorabolivar.com

Ilustración 9: Artefacto Mapa de Ciudadanos de Datos

En este artefacto nos dice la información oportuna, pero, para poder interpretarla de una manera rápida no es tan cómodo, por esta razón al practicante se le ocurrió una manera más legible visualmente para el mapeo de las personas, saber quién es el dueño, la influencia, etc.

Ilustración 10: Prototipo para el organigrama de emociones

Este es el prototipo de la información que va a tener cada persona, en este cuadro está la información del artefacto, pero de manera más visual, esto con el objetivo de hacer un organigrama, en este prototipo podemos ver la siguiente información, el nombre de la persona, el correo de contacto empresarial de la persona, también tenemos el cargo de esa

persona y la foto de dicha persona. Sin embargo, esa es la información básica, pero como en Gobierno de Datos es necesario tener más información importante para la oficina, esta información es el tipo de dominio de datos al que pertenecen, también podemos ver el interés que esa persona tiene con respecto a la oficina. Viendo solo el prototipo no da toda la información, porque este solo no dice el nivel de influencia de cada persona. De esta manera se ve el organigrama.

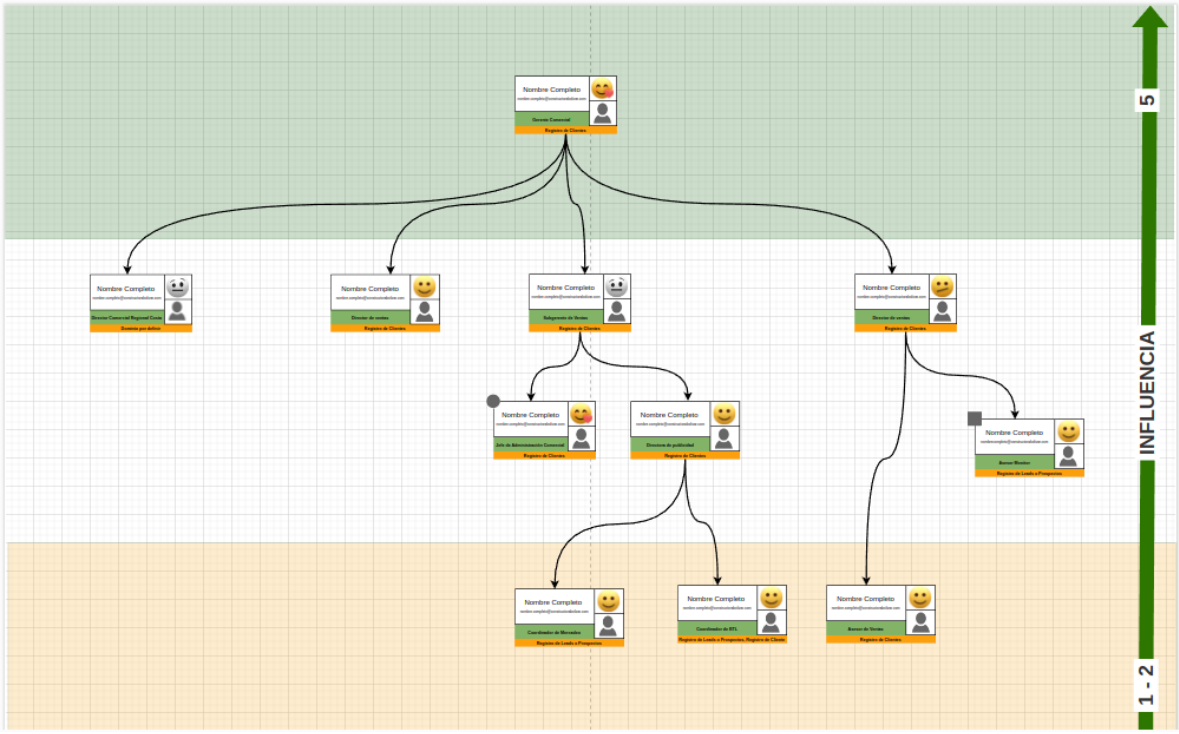

Ilustración 11: Organigrama de Emociones

Este organigrama dice muchas cosas importantes para la oficina de Gobierno de Datos, da la información de influencia, podemos observar la mayoría de las personas que tienen más influencias son las que están arriba del organigrama, y las que están más abajo generalmente tienen una influencia más baja. Este documento también dice a cuáles personas ya se entrevistaron y a cuáles no. Con esta vista podemos ver cuáles son las personas más interesadas en la oficina o las menos interesadas, esta vista nos ayuda a dimensionar estrategias para poder cambiar el estado de cada persona, una de las estrategias es enamorar a las personas que tienen más influencia y así hacer que ellos le cambien el estado a las personas que están más abajo. Acá también podemos ver quién es el dueño del dominio o los otros datos que se necesitan en Gobierno de Datos. Para poder entender más este organigrama es necesario saber sus estados o emociones identificadas de las personas, esta sería la documentación que se hizo para que los estados estén claros y las personas que le quieran

interpretar o darle mantenimiento a el organigrama de emociones lo hagan sin ningún inconveniente.

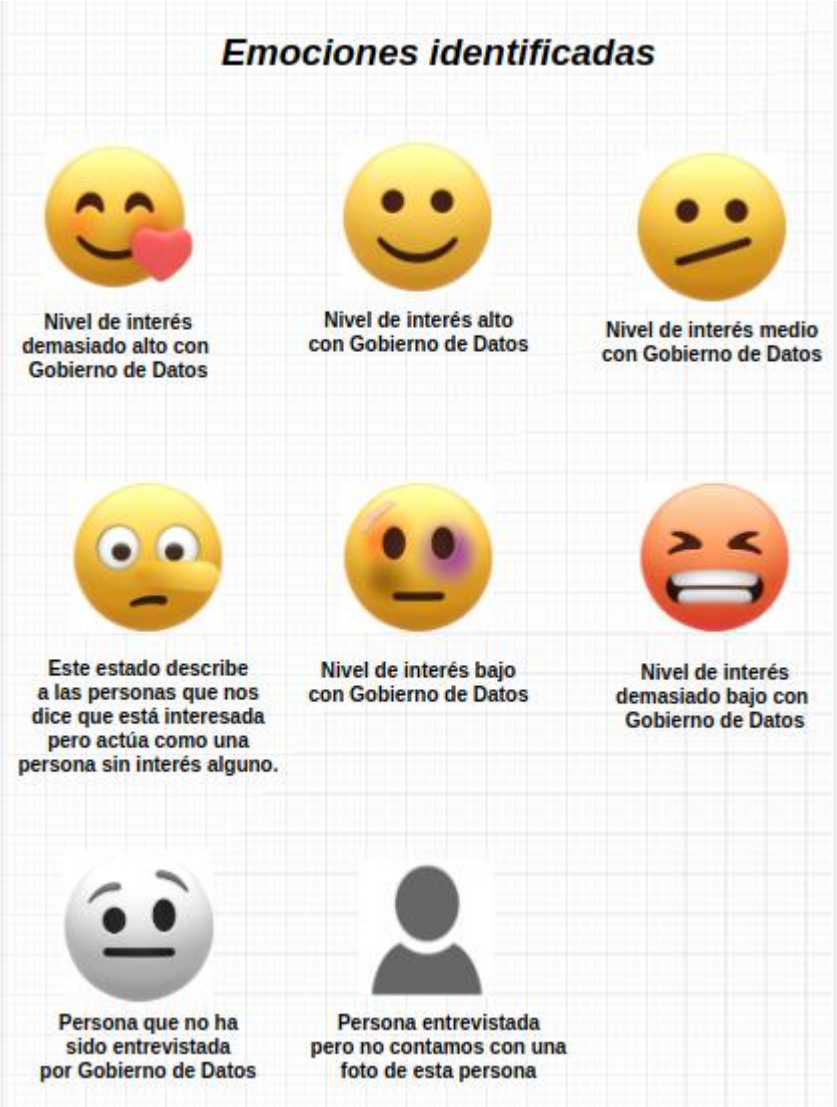

Ilustración 12: Documentación de las emociones

Todos estos artefactos mostrados o mencionados anteriormente, el practicante le daba mantenimientos, él acomodaba la información según lo que se había levantado. La mayoría de los artefactos era responsabilidad de todo el equipo, pero, algunos de ellos estos artefactos era cien por ciento responsabilidades del practicante.

Luego de eso se fueron sumando responsabilidades para todo el equipo, en la oficina estábamos haciendo muchas actividades manuales, por tal razón, necesitábamos de un

software que nos ayudara con la optimización y nos ayudara tener resultados más rápidos. Necesitábamos un software de gobernanza de datos, el objetivo principal con esta herramienta tener el linaje de los datos, entre otras cosas. En la mesa había varias propuestas, uno de los softwares propuestos era Collibra, finalmente no se contrató el servicio de este software porque era muy caro y demandaba tener una madurez de la calidad de los alta. Debido a esto, se contrató un servicio de Azure llamado Microsoft Purview, como en la compañía se usa Azure, era un punto positivo y la integración de las fuentes de nuestros era muy fácil. Gracias a esta situación, el jefe de la oficina le encargó al practicante a estar al mando de esta herramienta, tomó esa decisión por el desempeño que el practicante estaba teniendo dentro de su cargo. En esta etapa el practicante inició a estudiar temas de la nube, enfocado en Microsoft Purview. Fue acá donde aprendió sobre las ventajas y desventajas de Microsoft Purview, sin embargo, esto solo era la teoría y debía llevar todo ese conocimiento a la práctica. Dentro de este lapso el equipo de Gobierno de Datos hizo un spike donde el practicante era el que lideraba dicha reunión, fue donde socializó al equipo todo el conocimiento obtenido sobre el servicio. Debido a que este era el servicio que se iba a contratar y todo el equipo debía tener conocimiento de la herramienta.

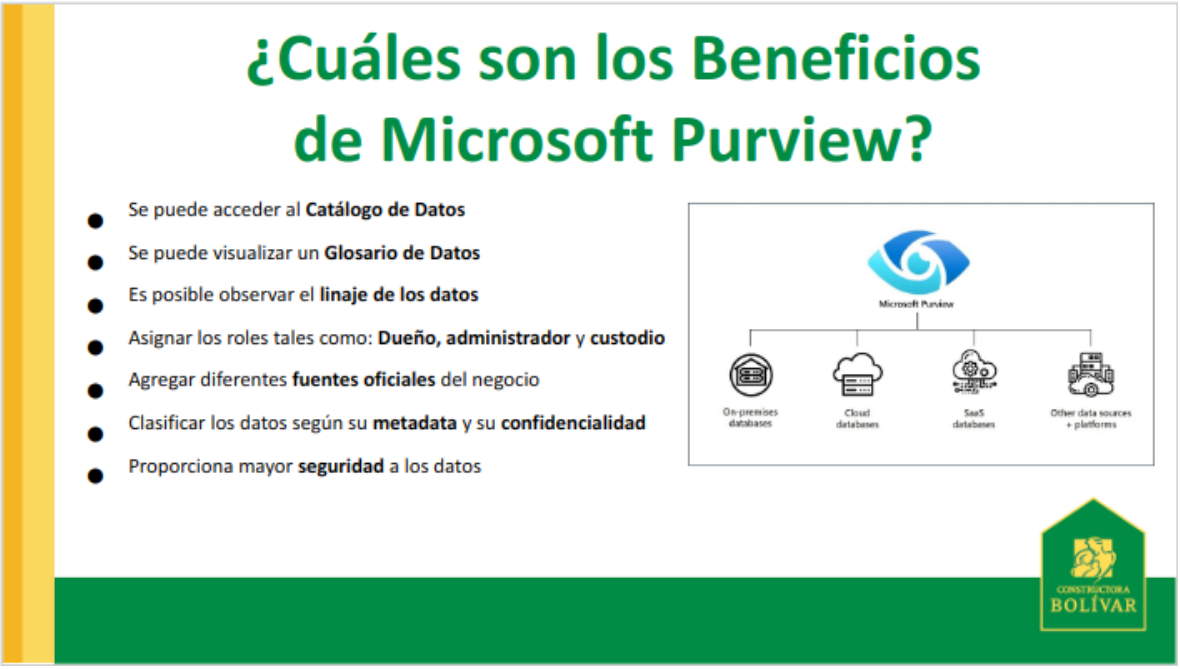

Ilustración 13: Slice de Spike

Este fue uno de los slice que el practicante expuso, donde adecuo la información de Microsoft Purview y sus beneficios a lo que el equipo estaba trabajando. Con esta socialización ya todo el grupo sabía en qué nos podía ayudar Microsoft Purview en nuestro trabajo día a día.

Después de eso, contrataron el servicio, y fue donde el grupo se enfrentó a varios retos. Había muchas cosas que nos ayudaban al objetivo de la oficina, era una lista importante de cosas, éramos capaz de agregar fuentes de datos, tanto On-Premises y en la nube, fue en reto total porque había que solicitar permiso de acceso a diferentes personas de diferentes áreas, como a las personas de infraestructura o a los arquitectos de datos. Todo esto se hizo de la mano de Accenture y del propio Microsoft, porque agregar las fuentes no dependía solo de nosotros y teníamos errores constantemente. Después de agregar todas las fuentes comenzamos a ver que ya no teníamos necesidad de agregar nosotros de manera manual los atributos, sino que el programa lo hacía de forma automática. En esta misma herramienta teníamos la posibilidad de agregar el glosario de negocio, con la diferencia que en la propia herramienta teníamos la opción de decir en qué estado estaba el concepto, si el concepto estaba en "borrador", o sea, que fue definido por la oficina de Gobierno de Datos, si el estado estaba en "certificado" significaba que este concepto ya fue discutido con el dueño del dominio y los diferentes administradores de los dominios. En esta plataforma hay estados como "expirado" lo que quiere decir que este concepto ya no se usa dentro de la compañía o que se usa con otro nombre.

Esta era una ventaja, pero, no se alineaba cien por ciento a nuestras tareas, porque la oficina no solo definía los conceptos del negocio sino que también definían los atributos de todos los dominios levantados. Entonces, para poder llevar esto en la plataforma se nos hacía cuesta arriba, ya que en el software no había una diferenciación entre definiciones del glosario y definiciones de los atributos. Este hallazgo fue expuesto al equipo de Accenture, donde me nos mencionaron algunas limitaciones de la herramienta y que no se podía trabajar de la manera como pretendíamos. Sin embargo, el equipo de Accenture nos brindó una solución alternativa, la cual fue implementada inmediatamente.

Utilizando a diario esta herramienta y llevando a cabo las peticiones de la oficina. Nosotros teníamos que saber cuál es el linaje de los datos, esto con el objetivo de saber dónde se ingresan los datos, cuáles son las áreas, gerencias y personas que los usan. Esto se hace con varios objetivos, uno de ellos es saber si tenemos alguna fuga de información y poder crear un plan de seguridad y no tener inseguridades de los datos. Además, tener el linaje de los datos es una ventaja porque de aquí podemos sacar conclusiones poderosas, una de ellas es saber de qué manera se están usando los datos en la compañía, saber si el uso de los mismo ha evolucionado, etc.

También podemos concluir cuales son los datos más importantes o usados en la compañía, con esto podemos saber si los datos de la compañía se están siendo aprovechado de la manera correcta. Teniendo esto claro podemos sacar conclusiones tales como: ¿A quién se le está brindando los datos?, ¿es correcto permitir que los datos los vea esta persona/área?, todo esto hay que preguntarse con el linaje de datos y si tenemos un aplicativo donde podamos ver esto

de una manera fácil. En la siguiente imagen podemos ver la manera como se ven los datos y su linaje en Microsoft Purview.

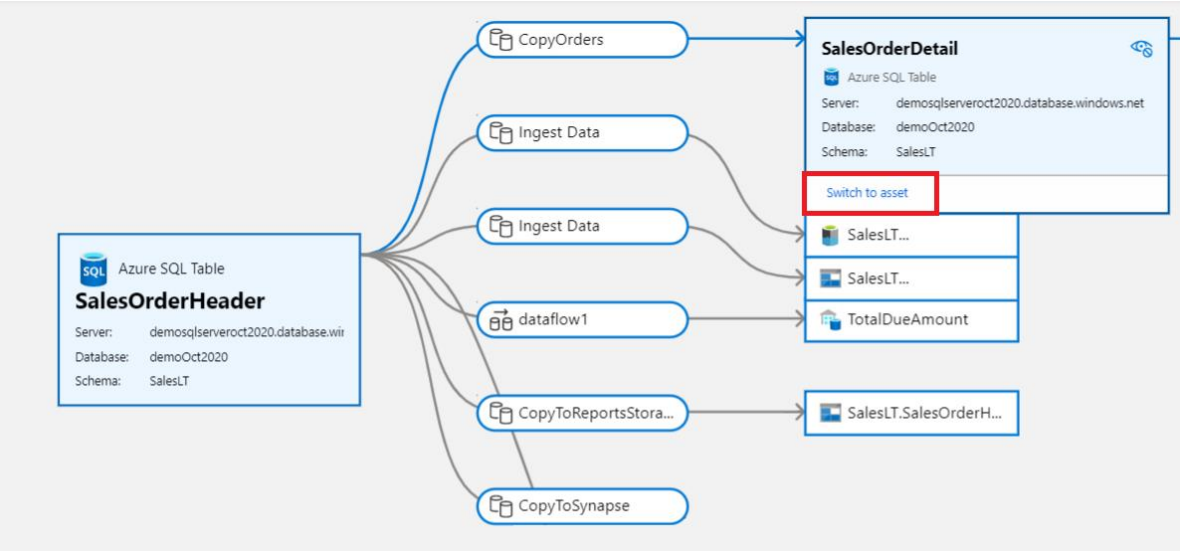

Ilustración 14: Linaje de Datos

Este linaje lo hace de manera automática la herramienta cuando se escanea las fuentes de información, eso importante porque podemos tener el linaje de manera actualizada, esta actualización no es en tiempo real, las actualizaciones dependen de las veces que se escanee las fuentes de datos, sin embargo, es ventaja porque de otra forma se tendría que hacer a mano alzada y los datos como están en cambio constante no los tendríamos actualizados. Otras de las ventajas es que en este aplicativo los metadatos se actualizan y se obtienen de forma automática, a diferencia del diccionario de datos que es manual y la información puede que no sea la correcta por los cambios de los datos en sí. Todas estas son ventajas, pero el software también tiene sus puntos negativos, uno de ellos es a la hora de obtener el linaje de los datos, cierto que este linaje se crea de manera automática, pero, esto depende de la tecnología que use los datos. ¿Qué quiere decir?, quiere decir que, si los datos se están trabajando en un software no soportado por Microsoft Purview, el linaje de los datos no se dibujará. Por esta razón es necesario incluirlos de manera manual, en el software se puede hacer de manera manual, pero hay algunos que no se proyectan. La otra manera es incluirlo por medio de la API con el lenguaje de programación Python. Para poder incluir el linaje de esta manera fue necesario desarrollar ese código. Como estábamos en proceso de aprendizaje de esta herramienta y no teníamos conocimiento de esta desventaja, hablamos con el equipo de Microsoft y nos comentó de la posibilidad de incluirlo por medio del código de Python, sin embargo, no era tan fácil como lo planteaba Microsoft. Por eso el equipo de Gobierno de Datos se puso manos a la obra para poder hacer el linaje, gracias a la guía de Accenture se

encontró la información. Fue aquí donde se comenzó a utilizar un código guía de GitHub y otra información que estaba en un canal de inglés en YouTube.

Para ejecutar este código es necesario hacer un par de pasos, primero es necesario crear una plantilla en Excel, esta plantilla se hace con un comando en Visual Studio Code. Ya teniendo esta plantilla era hora de hacer la conexión de Azure y el recurso destinado de Microsoft Purview, el código se ve la siguiente manera:

```
#Primer paso
from azure.identity import AzureCliCredential

from pyapacheatlas.core import PurviewClient

cred = AzureCliCredential()

# Create a client to connect to your service.
client = PurviewClient(
    account_name = "PURDataplatform",
    authentication = cred
)
```

Ilustración 15: Código de Python

Después era necesario autenticar el servicio, donde se le pasan los datos necesarios, tales como: Tenant_id, cliente_id_ y el client_secret. Con esto se valida el servicio principal, esta información se obtiene de Azure. En este mismo paso también es necesario hacer la autenticación y confirmar que sí se tiene acceso a los recursos que se están intentando acceder.

Ya teniendo estas configuraciones había que conectar la plantilla de Excel a Azure, acá por medio de código se puede crear todos los recursos sin necesidad de la interfaz. Podemos actualizar los recursos, crear nuevos atributos, crear los conceptos y también el linaje de datos. En este último era el que nos importaba hacer, el código que se encargaba de esta tarea es el siguiente.

```python
#Codigo para agregar nuevos recursos en Purview
from pyapacheatlas.readers import ExcelConfiguration, ExcelReader

ec = ExcelConfiguration()
reader = ExcelReader(ec)

processes = reader.parse_update_lineage("./my_file_path_nuevo.xlsx")
result = client.upload_entities(processes)
#print(json.drump(result, indent=2))
#print(result)
```

Ilustración 16: Código para actualizar el linaje de datos

Acá podemos ver que este código quiere actualizar el linaje, aunque para hacer esto no solo se hace la parte del código, en la plantilla se agregan los id de los recursos, se le dice quién es el transmisor o padre y quién es el receptor. Acá solo se está leyendo la plantilla y ejecutando todo lo que se agregó allá. Es importante entender esto porque agregar los recursos en la plantilla tiene su orden y no se puede hacer en cualquier forma, todo porque puede dar error. Ya teniendo este código y las funcionalidades podíamos hacer la actualización del linaje, otra cosa que hay que tener presente es que para hacer esto es necesario tener claro las conexiones que tiene cada recurso, porque como es manual si se dibuja mal el programa no va a hacer sugerencias, solo colocará las opciones o los valores que tengamos, eso en sí es una desventaja, ya que pasamos de tener linaje fidedigno y bien estructurado a tener algunos que quedan a conciencia de la persona que va a actualizar el linaje. La plantilla que se utiliza para dibujar el linaje de datos en Microsoft Purview se ve de la siguiente manera:

	A	B	C	D	E	F	G
1	Target typeName	Target qualifiedName	Source typeName	Source qualifiedName	Process name	Process qualifiedName	Process typeName
2	azure_datalake_gen2_	https://stcbdataplatformdev.dfs	azure_datalake_gen2_	https://stcbdataplatformdev.c	my_process	staging_bronze	Process
3							

Ilustración 17: Plntilla de Excel

Todo este levantamiento y entendimiento de la herramienta fue presentada al gerente y al director del área, en esta reunión se expuso toda la información que teníamos del sistema, ya quedaba en la decisión de ellos si contratar de manera permanente este software. A la conclusión que se llegó es que sí se iba a contratar a pesar de todas las limitaciones porque era una buena herramienta para empezar y que se ajustaba al presupuesto que se estaba dispuesto a pagar hasta el momento. Sin embargo, la contratación no iba a ser inmediata, porque había prioridades en el área que nos iba a consumir la mayoría del tiempo, en un futuro cuando no tengamos esas situaciones prioritarias ahí sí podríamos obtener el servicio.

Ya dejado cerrado el tema de Microsoft Purview, había que dar la mayor importancia a otras tareas pendientes del área. Acá nos centramos en hacer lo posible para crear los índices de

calidad de calidad. La visión era crear un dashboard donde nos mostrara los índices de los datos, para lograr este objetivo fue necesario hacer unos procesos en la compañía. Primero tuvimos que hablar con el equipo de BI, este equipo se compone por dos áreas, el área de requerimiento y los creadores de dashboard. Las primeras reuniones eran con el equipo de requerimiento, donde se nos preguntó el objetivo del tablero, quiénes lo iban a usar. Todo eran preguntas generales con el fin de determinar la prioridad y la importancia del tablero. Acá compartimos las reglas de calidad que se iban a evaluar y con ellos los atributos que queríamos mostrar en cada vista.

Fue acá donde se comenzó a hacer la ingesta en Data Factory de las tablas que se iban a usar, ya que todo esto se iba a trabajar en la nube, después de hacer la ingesta y guardar todos estos datos en Azure Data Lake House era necesario hacer todas las condiciones para satisfacer todas estas reglas. El desarrollo se hizo en Azure Databricks, primero el área de Ingeniería de datos no estaba ayudando para lograr hacerlo. Sin embargo, por cuestión personal nos quitaron las personas que no estaban ayudando y el practicante tomó el rol para que se pudiera continuar con el objetivo de hacer el dashboard.

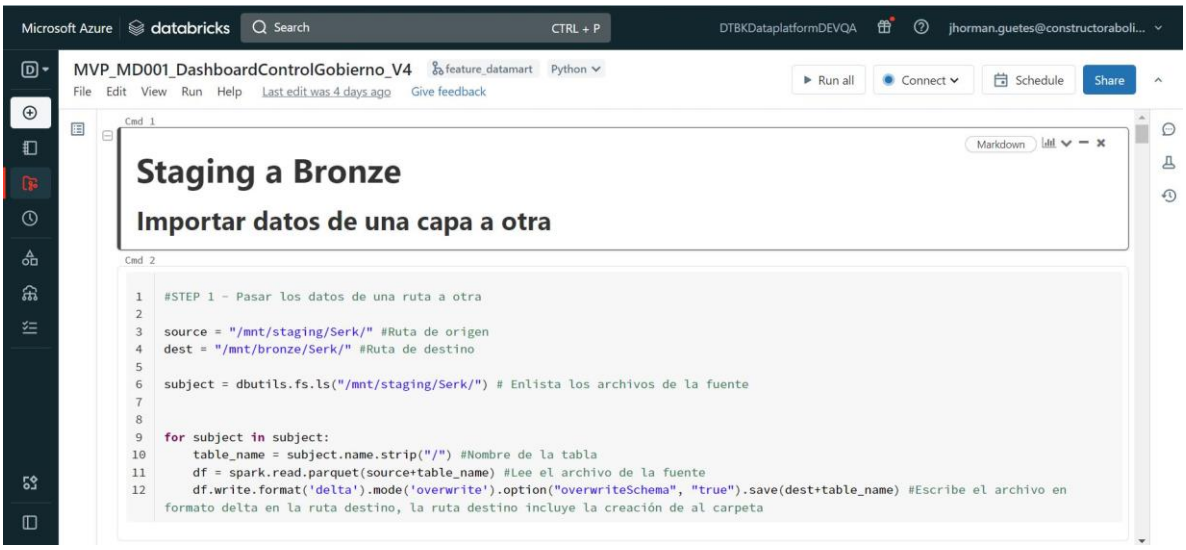

Ilustración 18: Copia de datos en capas

Acá podemos ver que los datos se están copiando de una capa a otra, porque es importante tener en una capa los datos en bruto, sin transformaciones. Ya en la siguiente capa lo que había que hacer era las transformaciones fuertes donde cumplan todas las condiciones ya planteadas. Cada notebook era responsable de un dominio de datos, el que estoy mostrando es solo para modificar los datos del dominio Registro de Clientes, ya que era el único tan avanzado para hacerle los índices de calidad.

Ilustración 19: Código de las dimensiones de calidad

En esta ilustración podemos ver cómo se ven las condiciones en Databricks, en este caso solo se está mostrando los campos que hasta ese momento se había agregado, en la dimensión de validez. En este caso tenemos atributos tales como correo electrónico. Todo esto se hacía con Python, una librería de este lenguaje llamado PySpark, este trabajaba con la tecnología de Apache Spark porque está tecnología tiene un buen rendimiento para trabajar volúmenes de datos grandes. Después de hacer estas transformaciones era necesario pasarlo a la capa donde se iba a disponibilizar, sin embargo, el equipo de BI que iba a hacer el tablero no tenían acceso a Azure Databricks, por tal razón, era necesario tomar esos datos y hacer una consulta en la herramienta de Azure Synapse, con el objetivo de que el equipo pudiera hacer lo suyo en Power BI. El resultado de todo este trabajo es el siguiente:

Ilustración 20: Tablero de Métricas de Calidad

En este tablero tenemos la siguiente información con diferentes filtros, podemos ver la calidad de datos por dominio, tenemos la información de la calidad por fuentes, el objetivo principal era medir la calidad de las tres fuentes más usadas de la compañía, cada una con sus complejidades. Podemos ver la calidad por atributo y el historial de cada medición que se ha hecho a lo largo de la compañía. Ya teniendo esta información en esta forma se debía implementar un plan de remediación, hay varias formas para hacer este plan de remediación, uno de ellos era hacer por medio de código en Databricks, cosa que estaba en la responsabilidad del practicante, esto se hacía con el fin de que los datos de la compañía tengan mejores datos. Otro de los métodos era buscar donde estaba flaqueando los datos, si se estaban ingresando mal a las fuentes o si directamente los datos se estaban manipulando después de ser ingresados. Todas estas bifurcaciones y posibilidades las tenemos que plantear con el dueño del dominio, hacer comités donde se iba a determinar cuál era la forma óptima. El papel de Gobierno de Datos era apoyar a las personas en este proceso, porque ellos podían conocer muchos los datos, pero muchas de las soluciones no sabían cómo implementarla y con el conocimiento de Gobierno de Datos podíamos ayudarles a tener la solución correcta. Como todo en esta empresa no debíamos tomar las decisiones de forma independiente por esa razón se socializaba a los gerentes los adelantos y las soluciones planteadas, para que ellos fueran esos impulsores o sponsors de nuestro trabajo. Con nuestras propuestas y trabajo estaban contentos porque la calidad de los datos no estaba en la mejor forma.

Ya finalizando las prácticas profesionales surgió un tema urgente dentro de la compañía, el tema que surgió era hacer un datamart, pero en el área de ingeniería de datos no tenía las manos suficientes para hacer el datamart. Debido al desempeño que traía el practicante su

jefe inmediato le dio la oportunidad de estar directamente relacionado con ese trabajo donde yo tuvo en la importancia y la relevancia en muchas reuniones. Por ejemplo, hizo aportaciones en el modelo dimensional en el que se estaba trabajando. También estuvo ayudando en la repartición de dimensiones que se iban a trabajar, donde se le asignaron facts y dimensiones que eran responsabilidad del practicante.

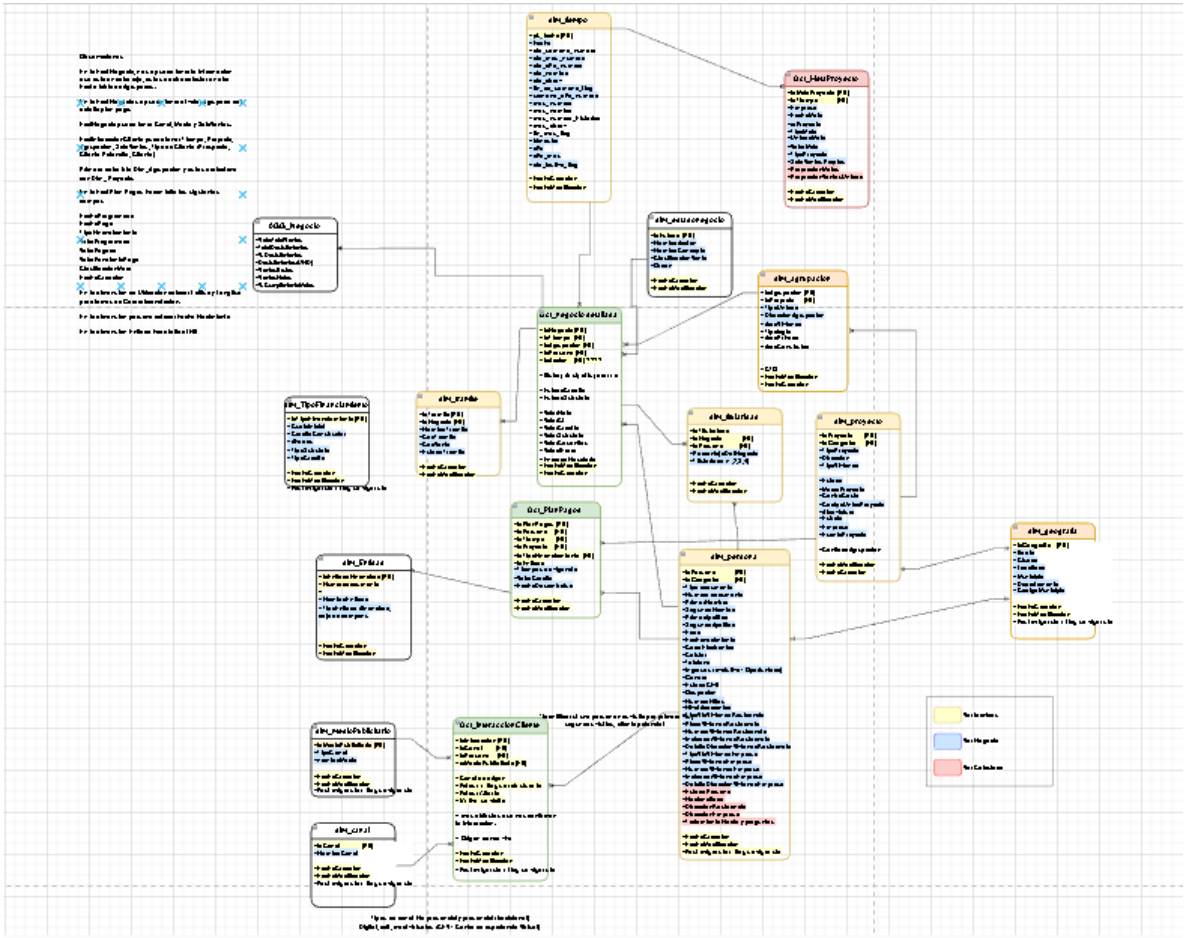

Ilustración 21: Modelo dimensional

En la imagen anterior podemos ver el modelo dimensional con el que se estaba trabajando para la creación del datamart, en este mismo modelo tenemos la posibilidad de ver cada una de las entidades, las dimensiones y los fats. Estas entidades incluyen los atributos y dichos atributos eran los que se debían mapear para hacer la ingesta en Data Factory y luego si hacer las transformaciones necesarias en Databricks, esto con el fin de disponibilizar la información al área correspondiente. La responsabilidad del practicante era mapear y buscar en las bases de datos los diferentes atributos que tenía cada entidad de este datamart, de tal modo que el practicante estaba en dos áreas, lo que quiere decir no estaba a cien por ciento en la oficina

de Gobierno de Datos sino que también estaba en la parte de Ingeniería de Datos, fue de esta manera como comenzó a asistir a dos daily y también comencé a trabajar en las dos áreas de diferentes. Fue ahí como llegó a demostrar todo el conocimiento y todas mis capacidades que tenía. Dentro de la empresa cuando se estaba haciendo el modelo dimensional había algunas cosas en las cuales el practicante intervenía porque consideraba que se podían hacer de una mejor manera, fue acá donde propuso algunas ideas, muchas de ellas fueron tomadas en cuenta y se implementaron. De esta manera fue como se ganó un espacio dentro del área de Ingeniería de Datos.

Disponibilizar la información del datamart era prioridad porque había algunos errores de calidad, también había de una manera u otros riesgos porque las áreas estaban tomando la información de cualquier lado y no estaba tomando dicha información de una fuente confiable o de la fuente de verdad. En la construcción del datamart Gobierno de Datos debía estar presente por muchas razones, la primera de ella es que la esa información debía estar segura y que las personas que usaran esa información la necesitaran. Era importante que esas personas que realmente necesitaban la información podrían acceder a ella de manera muy rápida y de manera muy eficiente. Por esa razón, también era importante que el área de porque se debía alinear y tener el control de la manipulación que se le estaba dando a los datos, con el fin de supervisar y validar las buenas prácticas que se estaban empleando, ya sea en la homologación de cada una de esas listas, debido a que Gobierno de Datos ya había hecho un levantamiento previo. Sin esta supervisión de la oficina era casi imposible cumplir el objetivo principal, el cual es organizar los datos y tener esa información ya estructurada con las buenas prácticas de DAMA DMBOK.

Las actividades anteriores mencionadas fueron las más importantes que desarrollaron dentro del proceso de las prácticas profesionales, cada una de ellas fue nombrado de manera cronológica sin embargo hay unas que estuvieron presentes a lo largo de las prácticas y fue el levantamiento de información y el entendimiento de las bases de datos. Estás dos fueron muy importantes porque en la oficina porque era necesario conocer el negocio de tal manera que pudiéramos controlar los datos de la compañía y esto solamente se podía hacer si uno tenía el conocimiento necesario para saber qué herramientas utilizar o cuáles eran los métodos correctos que ayudarán el objetivo era alinear las herramientas a la necesidad del negocio y no al revés porque cuando uno hace de que el negocio se alinea las herramientas lo que está haciendo es cambiar esa manera de operar de la empresa, ya que cada compañía tiene unas características muy puntuales las cuales no cualquier tipo de herramientas podría solventar. Entonces de esta manera nosotros éramos los que teníamos esa necesidad de saber de qué manera se estaba comportando el negocio y cuáles eran las mejores opciones dentro del mercado.

El otro proceso que hizo en todo el proceso de prácticas fue el entendimiento de las bases de datos, esta actividad también era importante porque ahí era donde se hacía el mapeo, era

donde se entendía de qué forma se conectaba una base de datos con otra, también de qué manera un atributo podría estar en diferentes tablas, cuál de los atributos era el que debía usar el negocio, a qué atributo hacerle medición de calidad y todo eso era un conocimiento constante.

Por esta razón la oficina estuvo envuelta en diferentes levantamientos de información tales como: El dominio de Registro del Cliente, el dominio de Registro de Leads y Prospectos, el dominio de Gestión de Trámites y todo esto con el fin de hacer la activación de cada uno de estos dominios. Todas estas tareas iban siempre con el objetivo de buscar los roles claves como el dueño del dominio, el administrador del dominio y el custodio, con el fin de tener esa estructura federada que se trazó a los inicios. Todas estas actividades fueron hechas a lo largo de las prácticas profesionales, esto iba de la mano con las entrevistas las cuáles constaban de preguntas claves y así saber los inconvenientes que tiene cada una de las áreas. Con este levantamiento se intentaba saber a los riesgos que estaban expuestos los datos de la compañía. Todas estas actividades se hicieron a lo largo de las prácticas profesionales.

8. CRONOGRAMA:

FASES	ACTIVIDAD	SEMANAS															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
FASE I	Entendimiento: ¿Qué es Gobierno de Datos?																
	Introducción: ¿Qué son los artefactos de Gobierno de Datos?																
	Entendimiento: La importancia de la calidad de datos																
FASE II	Mapeo de atributos del dominio de Registro de Clientes																
	Definición de las reglas de calidad de los atributos del dominio de Registro de Clientes																
	Levantamiento de información del dominio de Entregas																
	Mapeo de personas y gerencias involucradas en el dominio de Entregas																
	Identificar personas que pueden tener roles claves en el dominio de Entregas																

FASE III	Buscando información sobre Azure Purview																
	Socialización a mis compañeros sobre Azure Purview																
	Desarrollando código en Python para hacer el linaje de datos en Azure Purview																
	Socialización de Azure Purview con el gerente y el director																
FASE IV	Crear el SIR para hacer el dashboard de calidad de datos																
	Entendimiento de las herramientas de Azure para el desarrollo de la ETL																
	Desarrollo de la ETL de calidad de datos del dominio de registro del cliente																
	Socialización del dashboard de calidad de datos al gerente y al director del área																
FASE V	Ayuda a definir el modelo dimensional del datamart de venta																
	Mapeo los atributos del datamart																
FASE VI	Entendimiento del negocio																
	Entendimiento de las bases de datos																

9. CONCLUSIONES Y LÍNEAS FUTURAS

Podemos concluir que la oficina de Gobierno de Datos es importante dentro de las compañías, muchas de estas empresas necesitan esta oficina para tener ventaja competitiva o porque tienen regulaciones fuertes que las obligan a tener los datos de los clientes de la manera más correcta posible. También es necesario esta oficina porque en estos tiempos los datos son una materia prima muy importante que pocas veces se le da su valor correspondiente. En la actualidad se están manejando una cantidad de datos inmensa, lo que quiere decir que es necesario tener esos datos controlados, principalmente. Ya teniendo la información controlada y estructurada podemos hacer uso de ellos, con el fin de hacer reportes o tomar decisiones a través de ellos. Ya que de nada sirve tener muchos datos si esos datos no son confiables. ¿Por qué se necesita que los datos sean confiables?, porque a la hora de tomar decisiones con ellos podríamos estar tomando decisiones incorrectas por no tener datos exactos y veraces. Por esta razón podemos decir que la oficina de Gobierno de Datos es

importante, porque es la que intenta que estos datos sean lo más oportunos para las compañías. A pesar de ser muy importante, no es una oficina que se implementa de un día para otro, al contrario, para hacer una buena implementación de esta se necesitan años para lograr una buena madurez tanto de la oficina como de los datos. Esta es una de las razones por las que no se debe tratar la oficina de Gobierno de Datos como un proyecto, ya que esta no tiene fin, siempre está implementando reglas e intenta estructurar los datos de la manera correcta.

También podemos concluir que en el área de Gobierno de Datos para que funcione de una manera correcta es necesario que esté alineada con otras áreas tales como: Ingeniería de Datos, también debe de estar alineado con los arquitectos de datos, los responsables de la seguridad de la información. También se debe de estar alineados con los especialistas de calidad de datos, estos pueden ser ingenieros de datos o simplemente personas internas de la oficina. Si estas áreas y personas no están alineadas no se puede implementar bien la oficina.

Si la oficina sigue esta línea de operación en un futuro la compañía tendrá ventaja competitiva en comparación a las diferentes constructoras del país, además tendría menos probabilidad de ser denunciada por algún cliente debido a que en la actualidad no tratar bien los datos de los clientes hay probabilidades altas de que ellos mismos se quejen ante las autoridades y puedan acarrear sanciones económicas, entre otras cosas. Sin embargo, si se tiene el control de todos los datos que se tienen dentro de la compañía y cuáles se deben visualizar, se puede identificar de una manera más eficaz cuáles son los riesgos y de qué manera minimizarlos. Si se siguen abarcando los datos como Gobierno de Datos lo está haciendo podemos concluir que la compañía crecerá más y tomará las mejores decisiones a través de los datos.

10. BIBLIOGRAFÍA

Adobe. (Octubre de 2022). Recuperado el Diciembre de 2022, de Adobe: https://experienceleague.adobe.com/docs/experience-manager-65/assets/administer/metadata-concepts.html?lang=es

Amazon. (s.f.). Recuperado el Enero de 2023, de Amazon: https://aws.amazon.com/es/what-is/data-mart/

Amazon. (s.f.). Recuperado el Agosto de 2022, de Amazon: https://aws.amazon.com/es/what-is/python/

Amazon. (s.f.). Recuperado el Agosto de 2022, de Amazon: https://aws.amazon.com/es/what-is/sql/

Amazon. (s.f.). Recuperado el Agosto de 2022, de Amazon: https://aws.amazon.com/es/what-is/api/

Ariemetrics. (s.f.). Recuperado el Agosto de 2022, de Ariemetrics: https://www.arimetrics.com/glosario-digital/query

Arimetrics. (s.f.). Recuperado el Noviembre de 2022, de Arimetrics: https://www.arimetrics.com/glosario-digital/software

Computer Weekly. (s.f.). Recuperado el Noviembre de 2022, de Computer Weekly: https://www.computerweekly.com/es/definicion/Administrador-de-base-de-datos-o-DBA

DAMA DMBOK: Guia del conocimiento para la gestión de datos. (2017). El Comité Editorial Sanjay Shirude, Cathy Nolan, Emarie Pope. Recuperado el Noviembre de 2022

DAMA DMBOK: Guia del conocimiento para la gestión de datos. (2017). El Comité Editorial Sanjay Shirude, Cathy Nolan, Emarie Pope. Recuperado el Noviembre de 2022

DAMA DMBOK: Guia del conocimiento para la gestión de datos. (2017). El Comité Editorial Sanjay Shirude, Cath Nolan, Emarie Pope. Recuperado el Noviembre de 2022, de DAMA DMBOK: Guia del conocimiento para la gestión de datos.

Data Scientest. (s.f.). Recuperado el Noviembre de 2022, de Data Scientest: https://datascientest.com/es/pyspark-todo-sobre-la-biblioteca-python

Databricks. (s.f.). Recuperado el Noviembre de 2022, de Databricks: https://www.databricks.com/glossary/data-lakehouse

Digitus. (4 de Enero de 2023). Recuperado el Febrero de 2023, de Digitus: https://www.digitus.tv/que-es-dashboard/

EDTeam. (s.f.). Recuperado el Octubre de 2022, de EDTeam: https://ed.team/blog/estructuras-de-datos

EDX. (s.f.). Recuperado el Octubre de 2022, de EDX: https://www.edx.org/es/aprende/ingenieria-de-datos#:~:text=%C2%BFQu%C3%A9%20es%20la%20Ingenier%C3%ADa%20de,para%20el%20procesamiento%20de%20datos.

Google Cloud. (s.f.). Recuperado el Agosto de 2022, de Google Cloud: https://cloud.google.com/learn/what-is-data-governance?hl=es

Hernández, J. S. (s.f.). *Iessanvicente.* Recuperado el Febrero de 2023, de Iessanvicente: https://iessanvicente.com/colaboraciones/sqlserver.pdf

IBM. (s.f.). Recuperado el Octubre de 2022, de IBM: https://www.ibm.com/mx-es/analytics/hadoop/big-data-analytics

IBM. (s.f.). Recuperado el Octubre de 2022, de IBM: https://www.ibm.com/co-es/topics/business-intelligence

IBM. (s.f.). Recuperado el Octubre de 2022, de IBM: https://www.ibm.com/docs/es/cloud-paks/cp-data/3.5.0?topic=rules-quality

Microsoft. (s.f.). Recuperado el Agosto de 2022, de Microsoft: https://learn.microsoft.com/es-es/azure/data-factory/introduction

Microsoft. (s.f.). Recuperado el Enero de 2023, de Microsoft: https://azure.microsoft.com/es-es/solutions/data-lake/

Microsoft. (s.f.). Recuperado el Noviembre de 2022, de Microsoft: https://learn.microsoft.com/es-es/azure/databricks/introduction/

Microsoft. (s.f.). Recuperado el Noviembre de 2023, de Microsoft: https://learn.microsoft.com/es-es/azure/synapse-analytics/overview-what-is

Microsoft. (s.f.). Obtenido de Microsoft: https://powerbi.microsoft.com/es-es/what-is-power-bi/

Microsoft. (s.f.). Recuperado el Agosto de 2022, de Micorsoft: https://dynamics.microsoft.com/es-es/erp/define-erp/

Microsoft. (s.f.). Recuperado el Octubre de 2022, de Microsoft: https://learn.microsoft.com/es-es/purview/purview

Microsoft. (s.f.). Recuperado el Agosto de 2022, de Microsoft: https://visualstudio.microsoft.com/es/

Microsoft. (s.f.). Recuperado el Diciembre de 2022, de Microsoft: https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-azure/

Oracle. (s.f.). Recuperado el Agosto de 2022, de Oracle: https://www.oracle.com/mx/database/what-is-database/

Power Data. (2022). Recuperado el Octubre de 2022, de Power Data: https://www.powerdata.es/data-warehouse

Saltos, J. P. (2018). *LinkedIn*. Recuperado el Diciembre de 2022, de LinkedIn: https://www.linkedin.com/pulse/gobierno-de-informaci%C3%B3n-la-idea-realidad-jean-paul-saltos/?originalSubdomain=es

SAP. (s.f.). Recuperado el Enero de 2023, de SAP: https://www.sap.com/latinamerica/insights/what-is-data-governance.html

SAP. (s.f.). Recuperado el Octubre de 2022, de SAP: https://www.sap.com/latinamerica/insights/what-is-crm.html

Scrumio. (s.f.). Recuperado el Octubre de 2022, de Scrumio: https://www.scrumio.com/blog/que-es-un-spike/

Skyonline. (s.f.). Recuperado el Octubre de 2022, de Skyonline: https://www.skyonline.net/on-premise-vs-cloud-cual-es-la-mejor-opcion/

Sydle. (s.f.). Recuperado el Agosto de 2022, de Sydle: https://www.sydle.com/es/blog/data-driven-606c8a4e4b136c41e0e2c334

ANEXOS

GLOSARIO

Administradores del dominio

Los administradores del dominio son profesionales de negocios, expertos reconocidos en la materia, responsable de un subconjunto de datos. Trabajan con las partes interesadas para definir y controlar los datos. (DAMA DMBOK: Guia del conocimiento para la gestión de datos, 2017)

Apache Spark

Apache Spark es un Framework de código abierto desarrollado por el AMPLab de la UC Berkeley que permite procesar bases de datos masivas mediante computación distribuida, una técnica que consiste en explotar varias unidades de computación distribuidas en clusters en beneficio de un mismo proyecto para dividir el tiempo de ejecución de una consulta. Spark fue desarrollado en Scala y es mejor en su lenguaje nativo. (Data Scientest)

API

Las API son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos. Por ejemplo, el sistema de software del instituto de meteorología contiene datos meteorológicos diarios. La aplicación meteorológica de su teléfono "habla" con este sistema a través de las API y le muestra las actualizaciones meteorológicas diarias en su teléfono. (Amazon, s.f.)

Azure

La plataforma Azure está compuesta por más de 200 productos y servicios en la nube diseñados para ayudarle a dar vida a nuevas soluciones que permitan resolver las dificultades actuales y crear el futuro. Cree, ejecute y administre aplicaciones en varias nubes, en el entorno local y en el perímetro, con las herramientas y los marcos que prefiera. (Microsoft, s.f.)

Azure Databricks

La plataforma Azure Databricks Lakehouse proporciona un conjunto unificado de herramientas para crear, implementar, compartir y mantener soluciones de datos de nivel empresarial a escala. Azure Databricks se integra con el almacenamiento en la nube y la seguridad en su cuenta de nube, y administra e implementa la infraestructura en la nube en su nombre. (Microsoft)

Azure Data Factory

Azure Data Factory es un servicio en la nube administrado creado para estos complejos proyectos híbridos de extracción, transformación y carga (ETL), extracción, carga y transformación (ELT) e integración de datos. (Microsoft)

Azure Data Lake

Azure Data Lake incluye toda la funcionalidad necesaria para facilitar a los desarrolladores, los científicos de los datos y los analistas el almacenamiento de datos de cualquier tamaño, forma y velocidad, y para llevar a cabo todo tipo de procesamiento y análisis en diferentes plataformas y lenguajes. (Microsoft)

Azure Synapse Analytics

Azure Synapse es un servicio de análisis empresarial que acelera el tiempo necesario para obtener información de los sistemas de almacenamientos de datos y de macrodatos. Azure Synapse reúne lo mejor de las tecnologías SQL que se usan en el almacenamiento de datos empresariales, las tecnologías de Spark que se utilizan para macrodatos, Data Explorer para análisis de serie temporal y de registro, Pipelines para la integración de datos y ETL/ELT, y la integración profunda con otros servicios de Azure, como Power BI, CosmosDB y AzureML. (Microsoft)

Bases de datos

Una base de datos es una recopilación organizada de información o datos estructurados, que normalmente se almacena de forma electrónica en un sistema informático. Normalmente, una base de datos está controlada por un sistema de gestión de bases de datos (DBMS). En conjunto, los datos y el DBMS, junto con las aplicaciones asociadas a ellos, reciben el nombre de sistema de bases de datos, abreviado normalmente a simplemente base de datos. (Oracle)

Big Data

Big data es un término aplicado a los conjuntos de datos cuyo tamaño o tipo está más allá de la capacidad de la tradicional base de datos relacionales para capturar, gestionar y procesar los datos con baja latencia. Los big data tienen una o más de las características siguientes: alto volumen, alta velocidad o gran variedad. (IBM)

Business Intelligence

Business Intelligence (BI) es un tipo de software que se alimenta de datos de negocios y presenta reportes, paneles, tablas y gráficos de forma amigable para el usuario. Las herramientas de BI permiten a los usuarios de negocio acceder a diferentes tipos de datos: históricos y actuales, de terceros e internos, así como datos semiestructurados y datos no estructurados como las redes sociales. Los usuarios pueden analizar esta información para obtener insights sobre el rendimiento del negocio. (IBM)

Calidad de datos

El término calidad de datos se refiere tanto a las características asociadas con datos de alta calidad como a los procesos utilizados para medir o mejorar la calidad de éstos. Estos usos duales pueden ser confusos, por lo que ayuda a separarlos si se aclara qué constituye datos de alta calidad. (DAMA DMBOK: Guia del conocimiento para la gestión de datos, 2017)

CRM

CRM es la sigla en inglés para "gestión de relaciones con el cliente". Con un sistema de CRM, usted puede automatizar e integrar los datos y actividades relacionados con el cliente: ventas, marketing, servicio y e-commerce. (SAP)

Custodio de dominio

Los custodios del dominio son profesionales de TI que operan dentro de una de las Áreas de Conocimiento, como: especialistas de integración de datos, administradores de base de datos, especialistas en inteligencia de negocio, analistas de calidad de datos o administradores de Metadatos. (DAMA DMBOK: Guia del conocimiento para la gestión de datos, 2017)

DBA

Un administrador de base de datos (DBA) es el técnico de información responsable de dirigir o realizar todas las actividades relacionadas con el mantenimiento de un entorno de base de datos exitoso. Un DBA se asegura de que la base de datos de una organización y sus aplicaciones relacionadas funcionen de manera funcional y eficiente. (Computer Weekly, s.f.)

Dueño del dominio

Los Dueños de dominio son un Data Steward del Negocio, que tiene autoridad de aprobación para las decisiones sobre datos dentro de su dominio. (DAMA DMBOK: Guia del conocimiento para la gestión de datos, 2017)

Data Driven

Data Driven es un concepto estratégico que se puede aplicar a cualquier negocio. En la práctica, significa tomar decisiones basadas en el análisis y la interpretación de datos. Este término puede traducirse como gestión basada en datos. (Sydle, s.f.)

Dashboard

Un dashboard o panel de datos es una herramienta de gestión de la información que rastrea, analiza y muestra visualmente indicadores clave de rendimiento (KPI), métricas y puntos de datos clave para monitorear el estado de un negocio, departamento o proceso específico. Son personalizables. (Digitus, 2023)

Data Lake House

Un Data Lake House es una nueva arquitectura de gestión de datos abiertos que combina la flexibilidad, la rentabilidad y la escala de los lagos de datos con la gestión de datos y las transacciones ACID de los almacenes de datos, lo que permite la inteligencia empresarial (BI) y el aprendizaje automático (ML) en todos los datos. (Databricks)

Data Warehouse

Un Data Warehouse es un almacén electrónico donde generalmente una empresa u organización mantiene una gran cantidad de información. Los datos de un data warehouse deben almacenarse de forma segura, fiable, fácil de recuperar y fácil de administrar. (Power Data, 2022)

Data Mart

Un data mart es un sistema de almacenamiento de datos que contiene información específica de la unidad de negocio de una organización. Contiene una parte pequeña y específica de los datos que la empresa almacena en un sistema de almacenamiento más grande. Las empresas utilizan un data mart para analizar la información específica de cada departamento de manera más eficiente. Brinda datos resumidos que las partes interesadas clave pueden utilizar para tomar decisiones informadas rápidamente. (Amazon, s.f.)

Dimensión de calidad

Una dimensión de calidad de datos es una característica medible de los datos. El término dimensión se utiliza para hacer una analogía con la medición de objetos físicos (por ejemplo, longitud, ancho, altura). Las dimensiones de calidad de datos proporcionan un vocabulario para definir los requisitos de calidad de éstos. A partir de ahí, se pueden usar para definir los resultados de la evaluación inicial de la calidad de los datos, así como la medición continua. Para medir la calidad de los datos, una organización necesita establecer características que sean importantes para los procesos de negocio (que valen la pena medir) y medibles. Las dimensiones

proporcionan una base para reglas medibles, que deberían estar directamente conectadas a riesgos potenciales en procesos críticos. (DAMA DMBOK: Guia del conocimiento para la gestión de datos, 2017)

Dominio de datos

Un dominio de datos no es más que un conjunto de términos de negocio y atributos que se agrupan y pertenecen a un mismo concepto de negocio. Por ejemplo, el dominio de datos de clientes, productos, zonas geográficas, entre muchos otros. (Saltos, 2018)

Estructura de Datos

Las estructuras de datos en programación son diferentes formas de organizar información para manipular, buscar e insertar estos datos de manera eficiente. (EDTeam)

ERP

La planificación de recursos empresariales, también conocida como ERP, es un sistema que ayuda a automatizar y administrar los procesos empresariales de distintas áreas: finanzas, fabricación, venta al por menor, cadena de suministro, recursos humanos y operaciones. Los sistemas ERP desglosan los silos de datos e integran la información obtenida en los diversos departamentos, de esta forma, ayudan a los directivos a extraer conocimientos, optimizar operaciones y mejorar la toma de decisiones. (Microsoft)

Gobierno de Datos

El gobierno de datos abarca todas las medidas adoptadas para que los datos sean seguros, privados y precisos y, además, estén disponibles y se puedan usar. Incluye todas las acciones que deben realizar los usuarios, los procedimientos que deben seguir y la tecnología que utilizan durante todo el ciclo de vida de los datos. (Google Cloud, s.f.)

Gobernanza de Datos

La gobernanza de datos empresariales abarca las políticas y procedimientos que se implementan para garantizar que los datos de una organización sean precisos –y que se manejen correctamente cuando se ingresan, almacenan, manejan, acceden y eliminan–. Las responsabilidades de gobernanza de datos incluyen establecer la infraestructura y tecnología, configurar y mantener procesos y políticas, e identificar a las personas (o cargos) de una organización que tienen la autoridad y responsabilidad de gestionar y salvaguardar tipos específicos de datos. (SAP)

Ingeniería de Datos

La Ingeniería de Datos es una rama de las ciencias de la computación, que se ocupa del análisis de datos en grandes cantidades (Big Data), utilizando nuevas tecnologías de machine learning para el procesamiento de datos. (EDX)

SQL

El lenguaje de consulta estructurada (SQL) es un lenguaje de programación para almacenar y procesar información en una base de datos relacional. Una base de datos relacional almacena información en forma de tabla, con filas y columnas que representan diferentes atributos de datos y las diversas relaciones entre los valores de datos. Puede usar las instrucciones SQL para almacenar, actualizar, eliminar, buscar y recuperar información de la base de datos. También puede usar SQL para mantener y optimizar el rendimiento de la base de datos. (Amazon)

SQL Server

Microsoft SQL Server es un sistema de gestión de base de datos relacional (RDBMS) producido por Microsoft. Su principal lenguaje de consulta es Transact-SQL, una aplicación de las normas ANSI / ISO estándar Structured Query Language (SQL) utilizado por ambas Microsoft y Sybase. (Hernández)

Metadatos

Metadatos significa datos sobre los datos. A este respecto, los datos hacen referencia a su recurso digital, por ejemplo, una imagen. Los metadatos son esenciales para una administración eficiente de los recursos. (Adobe, 2022)

Microsoft Purview

Microsoft Purview es una familia de soluciones de cumplimiento, riesgos y gobernanza de datos que pueden ayudar a su organización a controlar, proteger y administrar todo el patrimonio de datos. (Microsoft, s.f.)

On-Premises

Conocido como on-premises/on-prem, este modelo informático consta de software descargado en un dispositivo físico que es propiedad de la organización, generalmente en formato conocido “licencia”. Las personas a menudo usan el término en referencia a los centros de datos de una organización, incluidos los servidores, los sistemas de red y los componentes del sistema de propiedad corporativa. Además, on-premises puede referirse a aplicaciones diseñadas internamente específicas para los casos de uso de la organización. (Skyonline, s.f.)

Power BI

Power BI es una plataforma unificada y escalable de inteligencia empresarial (BI) con funciones de autoservicio apta para grandes empresas. Conéctese a los datos,

visualícelos e incorpore sin problemas objetos visuales en las aplicaciones que usa todos los días. (Microsoft)

Python

Python es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine learning (ML). Los desarrolladores utilizan Python porque es eficiente y fácil de aprender, además de que se puede ejecutar en muchas plataformas diferentes. El software Python se puede descargar gratis, se integra bien a todos los tipos de sistemas y aumenta la velocidad del desarrollo. (Amazon)

PySpark

PySpark es una biblioteca que ofrece utilizarla con el lenguaje Python, manteniendo un rendimiento similar al de las implementaciones de Scala. Por lo tanto, Pyspark es una buena alternativa a la biblioteca Pandas cuando se quiere tratar con grandes conjuntos de datos que consumen demasiado tiempo de cálculo. (Data Scientest)

Query

Una query puede ser una palabra o cadenas de palabras (query strings) que se introducen en la caja de búsqueda de un motor de búsqueda para tener acceso a la información contenida en la web. Una query suele estar acompañada por algunos conjuntos de letras y otros operadores específicos, como por ejemplo booleanos, que facilitan su función. Esta consulta o query puede realizarse en una página que sea un motor de búsqueda en sí misma, como por ejemplo Google, Yahoo o Bing, o en un motor interno, on-site, siempre desde cualquier dispositivo. (Ariemetrics)

Reglas de calidad

Las reglas de calidad se ejecutan como parte del análisis de calidad de datos. La salida de una regla de calidad se muestra como una violación de regla en los resultados del análisis de puntuación de calidad de datos. Solo los datos de un único activo de datos se pueden enlazar con la definición de regla de datos y debe ser un activo de datos donde crea una regla de calidad. Puede enlazar datos desde un activo de datos diferente, pero entonces se guarda la regla de calidad en el activo de datos al que está enlazada. Las reglas de calidad no requieren un nombre. De forma predeterminada, tienen el nombre de la definición de regla de datos que se ha utilizado para enlazar los datos. (IBM)

Software

El software es un conjunto de reglas o programas que dan instrucciones a un ordenador para que realice tareas específicas. También se conoce como aplicaciones

de software, paquetes de software, herramientas de software y programas de software. El software puede utilizarse para gestionar datos, automatizar procesos y crear aplicaciones o productos informáticos. Su complejidad puede variar desde un simple programa de tratamiento de textos hasta complejos sistemas informáticos que controlan infraestructuras críticas en sectores como la sanidad y el transporte. (Arimetrics, s.f.)

Spike

El spike es un elemento del Product Backlog orientado a la investigación o experimentación, cuya finalidad es obtener el aprendizaje necesario para implementar la funcionalidad solicitada por el Product Owner o cliente. (Scrumio, s.f.)

Visual Studio Code

Visual Studio Code es un editor de código fuente ligero pero eficaz que se ejecuta en el escritorio y está disponible para Windows, macOS y Linux. Incluye compatibilidad integrada con JavaScript, TypeScript y Node.js, y cuenta con un amplio ecosistema de extensiones para otros lenguajes (como C++, C#, Java, Python, Go, .NET). (Microsoft, s.f.)

Número de ilustración	Descripción
Ilustración 1	En la ilustración número uno está el organigrama de la dirección de Big Data & Analytics, en esta se pueden ver las diferentes oficinas que pertenecen a la dirección.
Ilustración 2	En la ilustración número dos se puede ver la información sobre las ventajas y desventajas de los tipos de estructuras de Gobierno de Datos.
Ilustración 3	En la ilustración número tres se puede ver una de las inconsistencias de las bases de datos. Acá podemos ver que hay tablas en las bases de datos que están totalmente vacías y sin datos.
Ilustración 4	La ilustración número cuatro nos muestra el artefacto de Glosario de Negocio, acá se ve los diferentes campos que se diligencian para para llevar la información de los conceptos del negocio.
Ilustración 5	En la ilustración número cinco podemos ver el artefacto de Diccionario de Datos, este artefacto es donde se guarda la información de la metadata de los atributos.
Ilustración 6	En la ilustración número seis podemos ver el artefacto de Inventario de Atributos donde podemos ver el atributo y sus diferentes estados. También podemos ver en qué fuente existe o en cuál fuente no existe y si los atributos fueron evaluados por los DBA.

Ilustración 7	En la ilustración número siete está el artefacto de las Reglas de Calidad de Datos, en este artefacto podemos ver los atributos con sus respectivas reglas de calidad. En este artefacto se le asocian las dimensiones de calidad a los atributos y las reglas de calidad.
Ilustración 8	En la ilustración número ocho se puede ver el artefacto de Reglas de Calidad de Datos, acá podemos ver la lista de reglas de calidad de datos, estas son reglas estándar que se puede usar para atributos que tengan un comportamiento similar.
Ilustración 9	En la ilustración número nueve se muestra el artefacto Mapa de Ciudadanos de Datos, en esta podemos ver que dentro del artefacto se guardan todas las personas entrevistadas y la información de cada una de ella, tales como: Área, gerencia, información de contacto, rol en el dominio, etc.
Ilustración 10	En la ilustración número diez es el prototipo que se usa para el organigrama de emociones, acá podemos ver información de contacto y a qué dominios pertenece.
Ilustración 11	En la ilustración número once se puede ver el organigrama diseñado para ver a las personas entrevistadas, también podemos ver la influencia que tiene cada persona y el nivel de interés. En este organigrama podemos ver a las personas que se entrevistaron y a las personas que no se han entrevistado.
Ilustración 12	En la ilustración número doce se puede ver la documentación de las emociones del organigrama de emociones que es donde se llevan la información de las personas entrevistadas.
Ilustración 13	En la ilustración número trece se puede ver un slice realizada para el spike de la herramienta de Microsoft Purview.
Ilustración 14	En la ilustración número catorce se puede como se ve la gráfica del linaje de datos de en la herramienta de Gobernanza de Datos llamada Microsoft Purview.
Ilustración 15	En la ilustración número quince podemos ver el código de Python donde se agregan las credenciales y el nombre del recurso que se va a acceder. Esto es con el fin de manipular por medio de Python los recursos de Microsoft Purview.
Ilustración 16	En la ilustración número dieciséis podemos el código de Python hace que se actualice la gráfica del linaje de datos dentro de la herramienta de Microsoft Purview.
Ilustración 17	En la ilustración número diecisiete podemos ver la plantilla que se usa para agregar los diferentes recursos para actualizar el linaje de datos en la herramienta Microsoft Purview.

Ilustración 18	En la ilustración número dieciocho se puede ver el código que copia los datos de una capa de datos a la otra. Este código hace parte de un notebook en Databricks.
Ilustración 19	En la ilustración número diecinueve se puede ver el código que hace las reglas de calidad planteadas en el artefacto de reglas de calidad.
Ilustración 20	En la ilustración número veinte se puede observar el índice de calidad de datos, lo que quiere decir es que se puede ver el porcentaje de la calidad de los datos.
Ilustración 21	En la ilustración número veintiuno se puede ver el modelo dimensional para la creación del datamart de ventas de la compañía.